BASIC MATHEMATICS

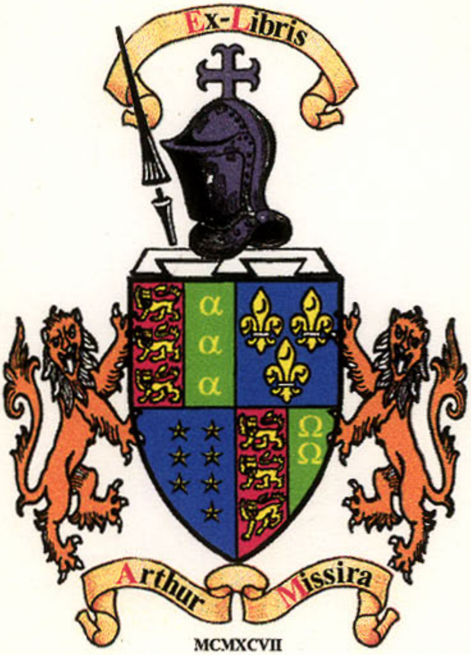
R.G.D. ALLEN

MACMILLAN

DEFINITIONS AND NOTATIONS

References are to sections of chapters

$\sum$	summation (1.7)
$n!$	n factorial (1.7)
$\binom{n}{r}$	binomial coefficient (1.7)
$\lvert a \rvert$	absolute value or modulus (1.7)
R	field of rational numbers (2.1, 2.2)
$R(\surd 2)$	adjunction, of $\surd 2$ to R (2.3)
R^*	field of real numbers (2.4)
GLB	greatest lower bound (LUB similarly) (2.4)
i	complex unit, $i^2=-1$ (2.5)
z	complex number $x+iy=r(\cos\theta+i\sin\theta)=re^{i\theta}$ (2.5, 12.7)
r, θ	absolute value and argument, of z (2.5)
C	field of complex numbers (2.5)
J^+	set of positive integers (natural numbers) (2.6)
J	integral domain of integers (2.6)
mod n	modulo n (2.7)
$F[x]$	polynomials over field F (3.3)
$F(x)$	rational fractions, adjunction of x to F (3.4)
ω	nth root of unity (3.8)
$\{a \mid a \text{ is } P\}$	set (4.1)
$\in$	belongs to (4.1)
$\subset$	proper subset of (4.1)
A'	complement, of set A (4.2)
$\cap, \cup$	intersection and union, of sets (4.2)
U, ϕ	universal and empty sets (4.2)
d, c	infinite cardinal numbers (4.7, 4.8)
$\sim p$	negation of p (not) (5.1)
$\wedge, \vee$	conjunction and disjunction (and, or) (5.1)
$\rightarrow$	implication, many-one mapping (5.1, 7.3)
$\leftrightarrow$	equivalence, one-one mapping (5.1, 7.3)
$P(a)$	probability, of statement a (5.5)
$P(a_1 \mid a_2)$	conditional probability, of a_1 given a_2 (5.6)
G	group, with identity e and a^{-1} inverse of a (6.2)
$r(A)$	operator; $sr(A)$ successive operators (r first, s second) (6.4)
F	field, with identities 0 and 1 (6.5)
$X \,.\, Y$	Cartesian product, of sets (7.1)
yRx	statement of a relation R (7.1)
$y=f(x)$	function (7.3)
$X \xrightarrow[f]{} Y$	mapping; $X \xrightarrow[T]{} Y$ transformation (7.3)
$\cong$	isomorphic (7.4)
$Z=f(z)$	function of a complex variable (7.6)
V	vector space (8.3)
$V_n(F)$	space of n-tuples over F (8.4)
$E_n(F)$	n-dimensional Euclidean space over F (8.4)
$(ABCD)$	cross-ratio $\dfrac{AB\,.\,CD}{AD\,.\,CB}$ (8.7)
$(1, \pm i, 0)$	circular points at infinity (8.8)
$[a, b]$	interval $a \leq x \leq b$ (9.3)
N	neighbourhood, of α (9.3)
$f^{-1}(x)$	inverse function (9.3)
$\mathop{\mathrm{Lim}}\limits_{n\to\infty} f(n)$	limit of $f(n)$ as n increases without bound (9.5)
$\mathop{\mathrm{Lim}}\limits_{x\to\alpha} f(x)$	limit of $f(x)$ as x approaches α (9.6)
$y', f'(x)$ $Dy, Df(x)$ $\dfrac{dy}{dx}, \dfrac{d}{dx}f(x)$	derivative of $y=f(x)$ (10.2)
$\int_a^b f(x)\,dx$	definite integral of $f(x)$, area (10.5)
$\int f(x)\,dx$ $D^{-1}f(x)$	indefinite integral of $f(x)$, anti-derivative (10.6, 10.8)
$D^n f(x)$	nth derivative, ($-n$)th integral, of $f(x)$ (10.8)
$\mathrm{Max}\, f(x)$	local maximum of $f(x)$ (minimum similarly) (11.2)
$\sum u_n$	infinite series (11.3)
$\sum a_n x^n$	power series, with radius of convergence r (11.6)
$\pi = 3{\cdot}14159\ldots$	Archimedes' constant (12.1, 12.5)
$e = 2{\cdot}71828\ldots$	Euler's constant, $\mathop{\mathrm{Lim}}\limits_{n\to\infty}\left(1+\dfrac{1}{n}\right)^n$ (12.1, 12.2)
e^x, $\exp x$	exponential function (12.2)
$\log x$	logarithmic function (12.3)
a^x, x^a	power functions (12.4)
$\cos x, \sin x$ $\tan x, \tan^{-1} x$	circular functions (12.5), trigonometric functions (12.7)
$\cosh x, \sinh x$	hyperbolic functions (12.6)
$\mathbf{A} = \lVert a_{rs} \rVert$	matrix (13.4)
$\mathbf{A}'$	transpose, of matrix (13.5)
$A = \lvert \mathbf{A} \rvert$	determinant, of matrix (13.6)
$\mathbf{A}^{-1}$	inverse, of matrix (13.6)
$\bar{y}(p)$	Laplace transform, of $y(t)$ (14.7)
$\bar{Y}(s)$	generating function, of Y_n (14.7)


Ex-Libris
Arthur Missira
MCMXCVII

Sr

=

BASIC MATHEMATICS

BASIC MATHEMATICS

BY

R. G. D. ALLEN

LONDON
MACMILLAN & CO LTD
NEW YORK · ST MARTIN'S PRESS
1962

MACMILLAN AND COMPANY LIMITED
St Martin's Street London WC 2
also Bombay Calcutta Madras Melbourne

THE MACMILLAN COMPANY OF CANADA LIMITED
Toronto

ST MARTIN'S PRESS INC
New York

PRINTED IN GREAT BRITAIN

PREFACE

'The great weakness of teaching in mathematics is that too much of it is concerned with training in mathematical jugglery and too little with education in mathematical ideas' (D. B. Welbourn).

THE NEW mathematics developed during the last 50–100 years are now well entrenched in honours courses at universities; they have scarcely made any impression at all on the teaching of mathematics in schools. Admittedly, a lag is to be expected here. It is only recently, both in the United States and in Europe, that the lag has been recognised as a serious one, as threatening the output of competent mathematicians at a time when the demand for them is rapidly increasing. The content of mathematical teaching, particularly at the critical phase of transition from school to university, is in urgent need of drastic change. With the situation deteriorating so obviously and so rapidly, several groups of mathematicians got together a few years ago in the United States to undertake urgent studies of mathematical teaching. The Committee on the Undergraduate Mathematical Program reported in 1955 and two volumes of *Universal Mathematics* had appeared by 1958. The 1957 Yearbook of the National Council of Teachers of Mathematics was devoted to *Insights into Modern Mathematics*. A Commission on Mathematics prepared a report which was published by the College Entrance Examination Board in 1959. A little later, the Organisation for European Economic Co-operation sponsored a seminar and a survey on the teaching of mathematics in schools in Europe and published their report, entitled *New Thinking in School Mathematics*, in May 1961. Discussions in Britain resulted in a report: *On Teaching Mathematics* (Editor: Bryan Thwaites), published by Pergamon Press in July 1961. The present text is designed primarily to assist in the promotion of the necessary reforms in the basic teaching of mathematics.

The point which I wish to stress above all others is that mathematics is an exciting, enjoyable and challenging study. To one who takes a broad view, mathematics not only has great range and power;

it also has essential unity and simplicity. To one who responds to a challenge, mathematics presents a vast field to explore, with boundaries both ill-defined and always expanding. To one who would have things done elegantly and economically, mathematics is a most satisfying discipline. Mathematicians aim at the best formulation, both as a rewarding exercise in itself, and to provide a foundation on which the super-structure of mathematical applications can be safely constructed. A preference for a beautifully neat and sound development, as opposed to a set of correct but messy proofs, is something much to be encouraged.

It is essential to penetrate the smoke screen laid down by so many mathematicians — their jargon and the devices they employ in 'solving problems'. Potential mathematicians need to appreciate the basic structure of mathematics, particularly with a view to the proper formulation of mathematical models in the sciences. They should be taught, not merely *how*, but *why* things are done in mathematics. Their education should be concentrated on mathematical ideas; dexterity in 'mathematical jugglery' in problem solving comes by experience. I have attempted to provide a short course on these lines, for a variety of people. They may be in their freshman year, beginning university courses in the natural, appied or social sciences. They may be in graduate schools, having discovered rather late that their subjects are treated mathematically. They may be teachers in schools, attempting to get across the basic mathematical ideas and to fire the imagination of young mathematicians.

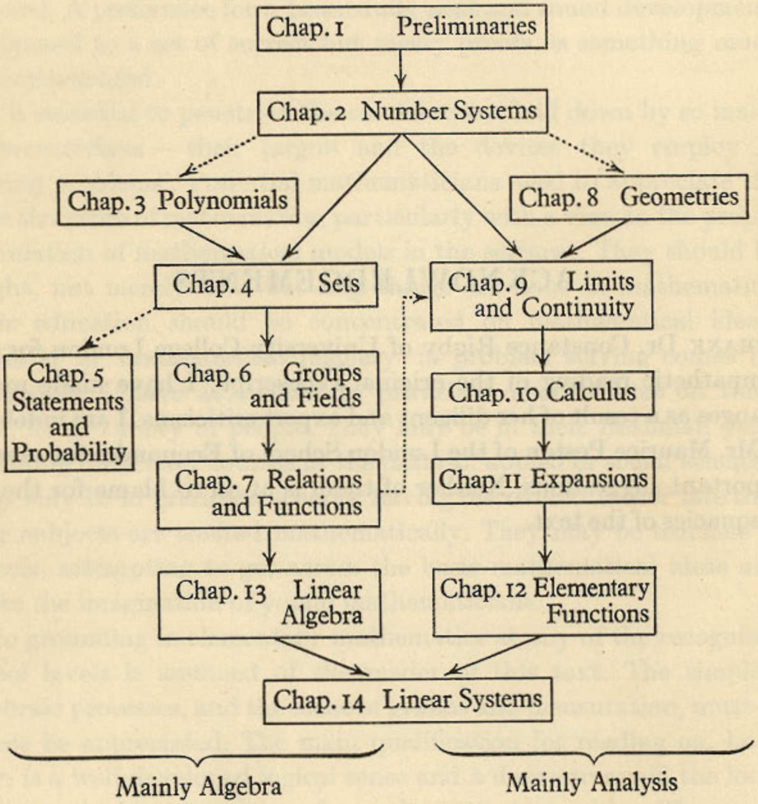
No grounding in elementary mathematics at any of the recognised school levels is assumed of the reader of this text. The simplest algebraic processes, and the ideas of graphs and mensuration, must of course be appreciated. The main qualification for reading on, however, is a well-developed logical sense and a desire to avoid the loose thinking which so often passes for mathematical exposition. The course presented is perhaps better followed under expert guidance rather than in private reading. A reader should have examples appropriate to his need suggested to him, and he should be given some indication of how basic mathematical ideas apply in his own field of study. Even so he must work hard, painfully sorting out his ideas; he is master of what he discovers for himself. I do not say that anyone who completes this course is thereby fully prepared to go on to such technical studies as matrix algebra or differential equations. I do maintain, however, that he will be much better prepared to do so.

London School of Economics R. G. D. Allen
July 1961

ACKNOWLEDGEMENTS

I thank Dr. Constance Rigby of University College London for her sympathetic reading of the original typescript; I have made many changes as a result of her diligent and expert criticisms. I am indebted to Mr. Maurice Peston of the London School of Economics for several important suggestions. Neither of them is at all to blame for the inadequacies of the text.

PLAN OF CHAPTERS



CONTENTS

9. Limits and Continuity

10. Calculus

11. Expansions

12. Elementary Functions

13. Linear Algebra

NOTE: * indicates exercises which are either difficult or involve new developments rather than illustrations of the text.

CHAPTER 1

PRELIMINARIES

'It might be as well to recall the professor who insisted that the essence of good teaching was always to tell the truth, but the whole truth only when students were mature enough to receive it.' *Report of the Commission on Mathematics* (College Entrance Examination Board, N.Y., 1959), Appendices, p. 64.

1.1. The traditional subjects of mathematics. In the teaching of mathematics in schools, it is customary to pass from very elementary work in arithmetic and mensuration to separate treatments of algebra and geometry, a little trigonometry being thrown in for good measure. Courses for more specialist students then develop into analysis, but there are still several compartments with separate labels: calculus, co-ordinate geometry, differential equations, and so forth.

To a schoolboy taught in this way, algebra must appear as an extension of arithmetic in which x stands for 'an unknown number' and in which more-or-less realistic problems are posed to provide exercises in 'finding x'. On the other hand, geometry is a drill in logical argument from an axiomatic basis through a long series of theorems, in the well-known Euclidean sequence, embellished by exercises in the form of riders on particular theorems. There may be some contact with reality at various stages, e.g. in dealing with volumes of solids or with trigonometric aspects of surveying.

It can be agreed that the 'whole truth' is not presentable at this stage in mathematical teaching; even so, is this the 'truth' about mathematics? It may be claimed that the traditional approach has worked well enough and that it has a certain logical convenience. But it is surely worth inspection.

At the practical level, algebra deals with the numerical aspects of things, and geometry with configurations in space, the two requiring different treatment. We are, however, somewhat shaken on discovering that, since geometry deals largely with distances, angles, areas and other measurements, the subject is numerical and wide open to algebraic treatment. We become almost schizophrenic in swinging from the abstract logical argument of 'pure geometry' to the algebra of 'co-ordinate geometry'. A good example of pure logic is to be found in projective geometry with its applications to perspective, to the deduction of properties of things from their shadows. But, when we come to investigate the properties of lines, circles, parabolas and other curves, we find it much easier to proceed in algebraic terms.

On teaching grounds, it may be said that mathematicians need to learn two things: to argue with strict logic from premises to conclusions; to acquire the right tricks for solving all kinds of problems as they arise. Algebra, as customarily taught, is a pretty good exercise in manipulation; geometry, on the Euclidean model, provides the necessary counter-weight, a training in logical reasoning from a postulational basis. Drop Euclid from the school curriculum and the teaching of mathematics becomes a matter of opening up a series of boxes of tricks; there would be none of the discipline of the axiomatic approach.

There is a flaw in the argument. Apart from minor difficulties, Euclidean geometry has one devastating defect as a complete and consistent axiomatic treatment. There is no postulate on the order of points in space, nothing to distinguish whether or not one point is between two other points. The difficulty is illustrated by the well-known 'proof' that every triangle is isosceles. In the triangle ABC, let AP be the bisector of the angle BAC and DQ the perpendicular bisector of the side BC, meeting in O, as in (i) of Fig. 1.1. Drop perpendiculars OE on the side AC and OF on AB. By this construction, OBD and OCD are congruent right-angled triangles and so are OAF and OAE. Hence:

$$OB = OC; \quad OF = OE; \quad AF = AE.$$

From the first two of these, OBF and OCE are congruent right-angled triangles so that

$$FB = EC$$

and $$AF = AE$$

already established. Adding:

$$AB = AC$$

and the triangle ABC is isosceles.

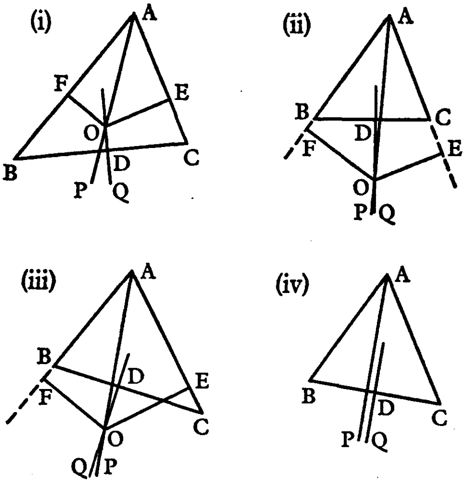


FIG. 1.1

The fallacy is in the last line. Can we add as indicated? There are three general possibilities and one special (or degenerate) case to consider, all illustrated in the figures:

(i) F is between A and B and E is between A and C.

Then: $$AB = AF + FB = AE + EC = AC.$$

(ii) F is not between A and B and E is not between A and C. Suppose F is beyond B and E beyond C as illustrated.

Then: $$AB = AF - FB = AE - EC = AC.$$

The same result follows in the only other possible situation, i.e. F and E beyond A.

(iii) F is not between A and B and E is between A and C (or conversely).

In the case illustrated:

$$AB = AF - FB = AE - EC \quad \text{but} \quad AC = AE + EC$$

and so $AB \neq AC$. A similar result holds in the converse case where F is between A and B and E is not between A and C.

(iv) AP and DQ are parallel or coincident, in which case O does not exist and the whole construction breaks down.

Which of these possibilities can hold? In fact, it can only be (iii) or (iv). If the triangle ABC is isosceles, then AP and DQ coincide, case (iv). If it is not isosceles, then the only possibility is (iii). But it is not possible to *prove* this by Euclidean geometry in default of any concept of the order of points on a line.

1.2. The axiomatic approach. Faced with this situation, we may try simply to plug the gap in Euclidean geometry. But is there not a similar problem of order for numbers and may it not be better to pay attention to the postulational basis of the number system and, indeed, of the whole of algebra? Why need we confine the axiomatic approach, even in elementary teaching, to the subject of geometry?

Mathematics is not a closed book. It has been growing vigorously over the centuries and there is still plenty of room left for development. In the past 50 years, it has become increasingly clear that the axiomatic basis of mathematics is not just an exercise for academic mathematicians and logicians. It serves to simplify, to unify and to generalise, to cast an illuminating light on the whole structure of mathematics and its applications. At the same time, mathematical formulations in the sciences — for example in such different fields as economics and physics — have become more abstract and sophisticated. It is now quite essential that the assumptions of mathematical models should be made precise and explicit, that the principles of model-building should be made clear.

The object of the following chapters is to expose the simple and uniform foundations of mathematics. In this we must be sure we have the truth and nothing but the truth — even if the whole truth sometimes eludes us if we are not to get bogged down in over-elaborate detail. It is apparent that, for appropriate application of mathematics as well as for our general satisfaction, it pays us to be careful and precise in formulation.* A word of warning: we can see the simplicity (as well as the uniformity and generality) of a sound axiomatic development once it is written down, but it is not easy to achieve. Advances are made by a nice combination of intuition and logic. Intution suggests what is to be established and the lines on which to lay out proofs. The best formulation, i.e. the most strict and economical, and the most revealing, is a matter for logical thought and experiment. The refinement and exposition of the concepts here described are the outcomes of the work of countless brilliant mathe-

* Hardy remarks in the preface to the third edition (1921) of his classic *Pure Mathematics*: 'It is curious to note how the direction of the criticisms I have had to meet has changed. I was too meticulous and pedantic for my pupils of fifteen years ago: I am altogether too popular for the Trinity scholar of today.'

maticians over the decades, indeed over the centuries. In approaching them we must be prepared to exercise our logical powers, but we should also leave plenty of scope for intuition.

We should forget the particular applications of mathematics with which we may be familiar and hence the particular division of subject matter into algebra, geometry, trigonometry, and so on. We are concerned with ideas which run through all mathematics. In tackling a problem, we may be used to starting: 'let x be any number' satisfying this or that condition; or 'let P be any point' on this or that locus. The basic idea here is: 'let x or P be any member of a set X', leading to a study of the set X, its specification, structure and properties. We must be prepared to have X as a set of entities of any kind. It may be a set of numbers, e.g. all positive integers or all real numbers. It may be a set of points in a plane, or the corresponding set of number pairs (the co-ordinates of a point). Or it may be a set of quite different entities, e.g. a set of operators or transformations. The vital matter is the structure of the set: how are the members related and how can they be combined one with another? It is a pleasant surprise to discover that sets of very different entities can have much the same structure.

There are other reasons for giving up the usual division of mathematics into subjects. Apart from the fact that we can sweep the lot together in an axiomatic approach, we can say that the compartments have never proved to be water-tight. Some simple equations of algebra, e.g. $x^2-2=0$ or $x^2+2=0$, cannot be solved within the system of rational numbers which characterises algebra; they need the real and complex number system of analysis. Further, algebraic expressions are often treated graphically, i.e. in terms of the geometric properties of curves. Conversely, much of geometry is best handled in algebraic form and with the aid of calculus. Such a simple geometric idea as the circumference or area of a circle involves a number π of a most sophisticated kind, a number which is not only 'irrational' but also 'transcendental', being the root of no polynomial equation. Again, the functions we first meet as trigonometric ratios appear in the most unlikely quarters, as sums of series and in the disguise of an area or an integral. And so on, back and forth between one subject and another.

Naturally, classification is one of the objects, and the delights, of mathematics. It just happens that classification by subject matter of application is not very useful in developing mathematical ideas. A better classification by far is the following. There is *finite mathematics,* dealing with finite sets, with stochastic processes and Markov chains, with vectors and matrices and with 'linear algebra' generally. There is the mathematics of the *countably infinite,* associated with the number systems of the integers and rationals, with particular reference to series and sequences. Then there is the most powerful mathematics of all: *mathematical analysis,* involving the continuum of real (and complex) numbers, leading through the idea of limits to the infinitesimal calculus.*

At this point, the geometers may raise a howl. The ideas and methods of geometry seem to be swallowed up by algebra and analysis. And so they are, once we think of a locus as a set of points, and choose to represent points in a plane by number pairs. On the other hand, we do find it convenient and helpful to keep the link between points and numbers always in mind. Geometric properties are translated into algebraic terms and, conversely, algebraic developments are illustrated visually in graphical or geometric terms. We can cater equally for those with an algebraic and those with a geometric turn of mind.

At all stages of a mathematical development, our constant aim must be to be both as precise and as general as we can. We need to be most clear on what we are doing in an axiomatic approach when we get down to fundamentals, when we start on such a major undertaking as the construction of the theory of sets or of the real number system. Clearly, no matter how deep we go in mathematics, we must have something to build upon, some straw for our brick-making. The basic discipline we must assume is the system of logical operations, framed in appropriate language and symbols. This is the logic of statements involving negation, disjunction, conjunction and implication, expressed by the words 'not', 'or', 'and' and 'if . . . then . . .' and written when convenient (as in 5.1 below) in terms of the symbols $\sim$, $\vee$, $\wedge$ and $\rightarrow$. On the axiomatic method, we proceed to add some-

* 'Calculus' is derived from the Latin: calx = stone. Modified by the diminutive 'ulus', it means 'small stone' as used in reckoning on an abacus. So calculus, or calculation, is any kind of reckoning. The adjectival qualification 'infinitesimal' is needed to indicate calculation with infinitesimal or continuous variation. Usually the 'infinitesimal calculus' is referred to simply as the 'calculus'.

thing new, something specific to the development in hand — and to be quite precise in formulating what is new. We first introduce certain primitive (undefined) concepts or relations and we follow these by defining other concepts or relations. Next, the concepts or relations are made subject to a system of axioms in the form of statements spelled out precisely to describe consistently and completely what properties we wish our new concepts to have. Only then are we ready to embark on the process of establishing further and consequent statements, i.e. to prove a sequence of theorems. The mathematical development, axiomatised in such a way, becomes a formal system or model.*

The ideas and methods involved in these chapters are essentially quite simple. It is true that they must be pursued with a certain determination, particularly when topics arise which are traditionally classified as 'advanced'. For example, having developed real numbers, we can go on quite naturally to complex numbers. It would be fatal to the whole approach to think of them as more difficult or as 'complex'; the label is indeed quite misleading. Looking back at the end, we shall see only two tough stretches — the definition of a real number and the consequent development of the idea of a limit — and we shall find only three or four results which are really difficult to establish. One is the fundamental theorem of algebra that a polynomial equation of the nth degree has precisely n roots. Another is the fundamental theorem of the calculus which establishes the inverse relationship between derivation and integration, and hence between the two apparently different applied processes of evaluating rates of change and areas.

1.3. Elementary algebra in terms of sets. As a preliminary canter over the field, we can attempt a re-interpretation of the elementary algebra of school text-books with emphasis on the basic concept of a set. The object is to make clear the logical foundation of algebra and to dispose of the notion that, in algebra as elsewhere in mathematics, the main thing is to acquire all the 'tricks of the trade'. We cannot dispense with tricks in mathematical work and some of the simpler

* A formal system developed strictly on the axiomatic method would be highly symbolised and a rather arid affair. Here we compromise in order to provide a general exposition which explains what is going on. It might be described as 'informal axiomatics'. See Church: *Introduction to Mathematical Logic* (Princeton, 1956).

ones are set out in the Appendix. A part of the skill of a practising mathematician lies in the box of tricks he has at his disposal; he needs to be expert in producing the right trick to solve the right problem. But it is much more important that he keeps always in mind the fact that everything he does is part of a broad and basic pattern, that the tricks he employs are not arbitrary and unrelated.

The way in which a 'formula' arises in mathematics can be shown by a simple example. Consider the assertion: the month of January has 31 days. Convert it into an open statement: the month of x has 31 days. What is x? We must be quite explicit. Here, we need to say that x stands for any one of the twelve months of the year. More precisely: x is any member of the set X comprising the twelve months. The statement is true of some members of X and false for others. The set X is the *replacement set* of the symbol x; we are entitled, in the statement, to replace x by any member of X we care to pick. More technically: the set X is the *domain* of the *variable* x. It is essential to know exactly what domain every variable has. In practice, the domain is often left to be understood from the context; it is always a good discipline to bring it out and to make it quite explicit.

The statement can be developed further: the month of x has y days. The domain X of the variable x is still the set of twelve months. What is y? Clearly y is another variable, but dependent in some way on the original variable x. It must also be a member of an appropriate set. Here, if leap years are ignored, y is a positive integer, one of a set of three $\{28, 30, 31\}$. Replace x by a particular member of its domain, and y is uniquely determined: either 28, or 30, or 31. The dependent variable y corresponds to a specific set Y, here $\{28, 30, 31\}$, and Y is called the *range* of y.

We can now appreciate what we have done in generalising our original assertion into the statement: y is the number of days in the month x. We have specified two sets: X comprising the twelve months of the year and Y consisting of three integers. The sets are related and the statement is a specification of how the linking is achieved. It gives a rule or a formula for going from a member of X to the corresponding member of Y.

In elementary algebra, we confine ourselves, rather strictly, to sets of numbers. We make generalised statements which we condense into formulae, of the well-known 'algebraic' type, involving certain numerical variables. Indeed, we tend to rush too quickly to the formula — too quickly to keep in mind precisely what we are doing. Too much emphasis is on the formula, too little on the sets being related. It is true that the formula is very interesting in itself, taking a variety of forms as illustrated below. The need remains, however, to specify what sets we deal with, before the precise nature of the formula connecting them can be appreciated.

1.4. Some examples

(i) Let x be any rational number, an integer or fraction (positive, negative or zero). Let y be the number obtained by taking twice the square of x, by subtracting x and then by subtracting 3. This rigmarole is condensed to the formula: $y=2x^2-x-3$. It is an *expression* of y in terms of x. Before we start to play around algebraically with the expression, we should pause to consider what sets we have. The domain X of the variable x is the set of all rational numbers, as opposed (for example) to the narrower set of integers or the wider one of real numbers. The values of y make up another set Y. The expression shows that y is always a rational number. What we do not yet know, and must find out, is whether the set Y comprises all or only some of the whole set of rationals. When this has been determined, the expression can be interpreted precisely as a formula for relating the set X to the set Y, for picking a member of the domain X and writing down the corresponding member of the range Y.

The domain X is the replacement set of the variable x. Pick any rational x from X and obtain the *value* of the expression by *substitution* of this x. For example, $y=2(\frac{1}{2})^2-(\frac{1}{2})-3=\frac{1}{2}-\frac{1}{2}-3=-3$ when $x=\frac{1}{2}$. In this way, a table of corresponding y's for selected x's is built up:

x	-2	-1	$-\frac{1}{2}$	0	$\frac{1}{4}$	$\frac{1}{2}$	1	$\frac{3}{2}$	2	3
y	7	0	-2	-3	$-\frac{25}{8}$	-3	-2	0	3	12

The table can be filled out by insertion of more and more entries. To study the nature of the expression $y=2x^2-x-3$, we find it helpful to draw a *graph* from the table of corresponding x's and y's and to observe the shape of the 'curve' shown. To plot the graph, take two axes Ox and Oy and an appropriate scale of measurement on each. It is usual to draw Ox horizontal and Oy vertical (Fig. 1.4). Given an x and associated y from the table, plot a point by going a distance x from O along Ox (to the right for positive x, to the left for negative x) and then by going a distance y parallel to Oy (upwards for positive y, downwards for negative y). The set of paired values of x and y in the table is then translated into a set of points on the graph. In this case, from Fig. 1.4, it appears that the plotted points lie on a smooth curve (confirmed later by theory) and a free-hand sketch of it can be included in the graph.

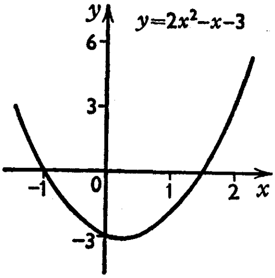


FIG. 1.4

We now have our answer to the question about the range Y of the expression. The graph indicates that, as x is allotted various rational values, so y takes all rational values except those less than a certain minimum. The smallest y is $-25/8$ when x is $1/4$. We seek to establish this algebraically and an appropriate way is easily found (Appendix A.5). So the expression $y=2x^2-x-3$ relates the set X of all rationals (the domain of x) to the set Y of all rationals not less than $-25/8$ (the range of y).

(ii) Consider the statement: think of a number (a positive integer), double it, add 4, divide the result by 2, subtract the number first thought of, and the answer is always 2. Write x for the optional number, with domain X as the set of all positive integers. The statement reduces to the formula:

$$\tfrac{1}{2}(2x+4)-x=2 \quad \text{for all } x \text{ in the set } X.$$

The formula here is an *identity*, true for all x's of the specified domain.* One way of expressing this fact is to write the expression $y=\frac{1}{2}(2x+4)-x$ defined for all x of X, and to show that the range of y is a set consisting of one item only, the integer 2.

As a common example of the use of identities, let x be any rational and consider the sequence of statements:

$$y=2x^2-x-3$$

$$2x^2-x-3=(x+1)(2x-3)$$

and so:

$$y=(x+1)(2x-3).$$

* In an identity the sign $=$ must be interpreted as 'identically equal'; it is sometimes written as $\equiv$.

The second of these is an identity. The other two are equivalent ways of writing the expression y. This is the familiar process of factorising.

(iii) Let x be any rational number and consider the statement that twice the square of x is equal to x plus 3. This may be true for some x in the domain X of all rationals, but it is certainly false for others. The statement can be written as the *equation*:

$$2x^2 = x + 3 \quad \text{or} \quad 2x^2 - x - 3 = 0$$

to be considered over the domain X of all rationals.

Write the expression $y = 2x^2 - x - 3$ and reverse the procedure of (i). Instead of substituting various x's and writing corresponding y's, we now ask: given that y is zero, what x's in the domain X will do? The graphical approach works again. From Fig. 1.4, we locate two x's which will do, the values -1 and $3/2$. In seeking to establish this algebraically, we find the factorisation of (ii) turns the trick: since $y = (x+1)(2x-3)$, it follows that $y = 0$ only when $x = -1$ or $x = 3/2$.

The achievement here is to start with the set X of all rationals and to narrow it down to a set of two rationals $\{-1, 3/2\}$ for each of which the equation $2x^2 - x - 3 = 0$ is valid. This narrower set of valid x's is the *solution set* of the equation. In writing $x = -1$ or $x = 3/2$, we have 'solved' the equation $2x^2 - x - 3 = 0$.

(iv) Vary the statement to read: twice the square of x is less than x plus 3. Again this is the kind of statement true for some and false for other x. The formula is now an *inequality*:

$$2x^2 < x + 3 \quad \text{or} \quad 2x^2 - x - 3 < 0$$

to be considered over the domain X of all rationals.

The same question arises: what x's in the domain X will do? In the graphical approach (Fig. 1.4), we seek those x's which make $y = 2x^2 - x - 3$ negative and for which the curve falls below the axis Ox. The answer is apparent: all x's in the interval between -1 and $3/2$. The algebraic proof is to be supplied: since $y = (x+1)(2x-3)$, it follows that $x < -1$ gives $y > 0$, $-1 < x < 3/2$ gives $y < 0$, and $x > 3/2$ gives $y > 0$ again. The set of valid x's is the *solution set* of the inequality. Once again, the formula (inequality here) enables us to cut down the set X of all rationals to a smaller set. In this case, the (smaller) solution set happens to be the set of all rationals x such that $-1 < x < 3/2$. We say that the inequality $2x^2 - x - 3 < 0$ holds for rational x in the interval $-1 < x < 3/2$.

1.5. Variables, constants and parameters. We have now a good idea of what we mean by a variable. A set X of items (usually numbers in algebra) is specified; a *variable* x is any member of the set X, and X is the replacement set or domain of the variable. A variable is a 'place-holder' for any member of its domain. It is not good enough to describe a variable as an 'unknown'. A variable is a member of a precisely specified set (its domain) and we know exactly what values we allow it to have.

The difficulty arises because of loose thinking about the solution of equations or inequalities. Suppose we are given some expression y in the variable x with domain X. We may seek those particular x's which make $y=0$ (an equation) or which give $y<0$ (an inequality). Some x's will do, others not. The result is the solution set, a narrowing down of the domain X to a smaller set. For example, if X is the set of all rationals, the solution set of the equation $2x^2-x-3=0$ is the set $\{-1, 3/2\}$ of two rationals and that of the inequality $2x^2-x-3<0$ is the set of rationals x such that $-1<x<3/2$. It is not an adequate description of this process to say that the variable x is an 'unknown' in the equation $2x^2-x-3=0$ and then to 'find' it as either -1 or $3/2$. The process in the use of the formula (equation and inequality alike) is one of cutting down the domain X to some smaller set, the solution set.

There is some lack of agreement on the terminology in the solution of equations. We know that, if we substitute either $x=-1$ or $x=3/2$ in the expression $y=2x^2-x-3$, then we get $y=0$. How can we describe this most conveniently? The value -1 or $3/2$ may be called a 'solution' or a 'root' of the equation $2x^2-x-3=0$. The same value may be called a 'root' or a 'zero' of the polynomial $y=2x^2-x-3$. The following is the terminology adopted here. The value $x=-1$ (or $x=3/2$) is a *zero* of the *expression* $y=2x^2-x-3$; on substitution we find $y=0$. The value $x=-1$ (or $x=3/2$) is a *root* of the *equation* $2x^2-x-3=0$; it is a value for which the equation is valid. The complete set, here $\{-1, 3/2\}$, of all roots is the *solution set* of the equation. The term *solution* is reserved for the process of finding the roots. In the illustration given, the expression happens to be a polynomial, and the equation a polynomial equation. The terms apply just as well to any expression $y=f(x)$ in a variable x with domain X comprising real numbers. It is useful to have two terms (zero and root) for what might appear to be the same thing. There is, in fact, a distinction which is worth making. An expression has zeros; an equation has roots. The link is that a zero of the expression $y=f(x)$ is also a root of the equation $f(x)=0$; each is a value of x which, on substitution, makes $f(x)$ vanish. The term to use depends on whether we have in mind the expression or the equation.*

Interest in algebra is concentrated on relations between variables, particularly when there is a dependent variable y with range Y given uniquely in terms of a variable x over a domain X. The dependence is shown by some algebraic expression such as $y=2x^2-x-3$. There are the two things to keep in mind. One is that the dependence is a linking of two sets, the domain X and the range Y. The other is that the dependence is expressed by means of some formula or other, and that there is a great variety of such formulae.

In a specified algebraic expression, in addition to the variable x, there are certain particular numbers called *constants*, all combined by the simple processes of algebra. If the expression is $y=2x^2-x-3$, the constants are 2, -1 and -3 and the only algebraic processes are addition and multiplication: $y=2\times x\times x+(-1)\times x+(-3)$. However, the notation is flexible enough to accommodate more general cases and to lead to more powerful methods of analysis. The expression $2x^2-x-3$ is recognised as just one instance of a whole class of quadratic polynomials with rational constants as coefficients. In this case the coefficients happen to be 2, -1, and -3. There are many other instances, for example $x^2-\frac{1}{2}x+\frac{1}{4}$ with coefficients 1, $-\frac{1}{2}$ and $\frac{1}{4}$. We cannot possibly list them all; we would like a notation enabling us to speak of *any* quadratic polynomial with rational coefficients. This is a simple matter, once we allow for coefficients which are *parameters*.† We write a general quadratic polynomial as $y=ax^2+bx+c$, where x is the variable and where a, b and c are

* Different terms are used by various writers. Two things are clear enough: a polynomial equation has roots, a function of a complex variable has zeros. The lack of agreement arises between these extremes. It is to be noticed that this use of 'root' is an extension of a simpler usage. The 'cube root' of a real number a is the single real value which, when raised to the third power, gives a. It is one root of the equation $x^3=a$ (in the present sense). There are also two other roots which are conjugate complex. It might be preferable to reserve the term 'root' for the ordinary concept of the nth root of a positive real number a, i.e. for the single positive real value satisfying $x^n=a$. We could then use 'zero' both for functions and for equations. But it is established terminology to speak of the roots of an equation.

† 'Parameter' is derived from the Greek: para = beyond, metria = measuring.

parameters. From one point of view, the parameters are constants, but not particularised. We then think of $y=ax^2+bx+c$ as one quadratic without specifying which one. From another point of view, the parameters are variable, when we switch from one quadratic to another or consider a whole class of quadratics.

An essential feature of the parametric form is that we must specify the replacement set of the parameters. The coefficients in $y=ax^2+bx+c$ are drawn from some specific set, which may be the same as or different from the sets of values of x and y. For example, the domain X may be all rational numbers (and the range Y some other set of rationals) while the parameters a, b and c may also come from the set of rationals, or they may be drawn from the set of integers. Hence, we have the qualification: $y=ax^2+bx+c$ is the general quadratic with rational coefficients, with integral coefficients, or whatever may be specified.

The parametric notation is one of great power. It enables us to deal with a class of expressions of the same general type, to establish properties valid for all expressions of the class. For example, it can be shown (Appendix A.5) that the class of quadratic polynomials $y=ax^2+bx+c$ with rational coefficients, and such that $a>0$, has the common property that each has a single minimum. The graph of each is of the form shown in Fig. 1.4. We can also locate the minimum; it is $y=-(b^2-4ac)/4a$ attained where $x=-b/2a$. Again we may seek the roots of the quadratic equation $ax^2+bx+c=0$, where a, b and c are rationals such that $b^2>4ac$. We have our answer (Appendix A.3): there are two roots $\{-b\pm\surd(b^2-4ac)\}/2a$. We can always proceed from the general to the particular, by specifying the values of the parameters. For example, put $a=2$, $b=-1$ and $c=-3$, and we find that $y=2x^2-x-3$ has the smallest value $-25/8$ when $x=\frac{1}{4}$, and that $2x^2-x-3=0$ has the two roots $\frac{1}{4}(1\pm5)$, i.e. -1 and $3/2$.

1.6. Unresolved problems in elementary algebra. It is perhaps inevitable that the treatment of algebra at elementary levels glosses over certain difficulties of a rather troublesome nature. It is a good exercise to bring these difficulties out into the open, even if they cannot be overcome immediately. Several of them are considered here.

(i) Consider the expression $y=a^x$ where x is a variable and a is a parameter. This is a power of a with a variable exponent x. What is an appropriate domain for x and what replacement set can be used for a? In the customary usage (Appendix A.1), a^x is a multiplied by itself x times when x is a positive integer, it stands for a qth root when x is a positive fraction p/q, and it is the reciprocal of the corresponding positive power when x is negative (integral or fractional). This is as far as elementary algebra goes. No meaning is attached to irrational powers such as $a^{\sqrt{2}}$ or a^{π}. Hence, in writing $y=a^x$, the domain of x cannot be wider than the set of rationals. Something more is needed, outside the scope of elementary algebra, if x is to be taken over the domain of all real numbers.

On the other hand, the parameter a can be drawn from the set of real numbers, rational or irrational: 2^{π} is not defined but π^2 certainly is. The difficulty is that a cannot be negative for certain values of x in a^x; for example, $(-64)^{1/3}=\sqrt[3]{(-64)}=-4$, but $(-64)^{\frac{1}{2}}=\sqrt{(-64)}$ cannot be written. Further, $a=0$ is possible in a^x for some x but not for others; for example $0^2=0^{\frac{1}{2}}=0$, but no meaning is attached either to 0^{-2} or to $0^{-\frac{1}{2}}$ since we cannot handle zero in the denominator. The most we can do, in elementary algebra, if we take $y=a^x$ over the domain of all rational x, is to limit a to the set of positive real numbers.

It is to be noticed, in particular, that:

$$a^{\frac{1}{2}}=\sqrt{a}=\textit{positive}\text{ square root of }\textit{positive }a.$$

No square root is written of a negative value. But if a is positive, then there is a positive square root $\sqrt{a}$, and also another or negative square root to be written $-\sqrt{a}$. So, if $x^2=a$, then $x=\sqrt{a}$ and $x=-\sqrt{a}$ are the two roots.

(ii) The consequence is that the concept of common logarithms, as used in ordinary numerical work, is seriously undermined (Appendix A.2). The notation of a logarithm as an inverse power looks innocent enough. In writing $y=\log_{10}x$ we mean that $x=10^y$. A common logarithm is the exponent in a power of 10. We have just seen that, as far as elementary algebra goes, we can write 10^y only when y is rational. There is no difficulty with some logarithms; for example, $\log_{10}0{\cdot}01=-2$ since $10^{-2}=0{\cdot}01$, and $\log_{10}\sqrt{10}=0{\cdot}5$ since $10^{0\cdot5}=\sqrt{10}$. But what of $\log_{10}2$ and many others? Tables of common logarithms give us a value of $\log_{10}2$, i.e. 0·3010 to four decimal places, 0·30103 to five decimal places, and so on. The implication is that

$$\log_{10}2=0{\cdot}30103\ \dots,$$

an irrational number to be approximated to as many decimal places as we wish. This would mean that $10^{0\cdot30103\ldots}=2$. And this is something we cannot write, at least in elementary algebra.* To justify completely the practical use of tables of logarithms requires more than elementary algebra can provide.

(iii) Consider the quadratic equation $ax^2+bx+c=0$ with coefficients which are real numbers ($a\neq0$). For the moment, leave open what domain we have in mind for x. The solution of the equation leads to:

$$x=\{-b\pm\sqrt{(b^2-4ac)}\}/2a. \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

We seem to have the very convenient result that every quadratic equation has two roots, one given by the + sign and the other by the − sign in (1). This is, however, more than the formula (1) can support. If the domain of x is the set of all rationals, then we recognise only rational roots. Hence, if b^2-4ac is a perfect square, the quadratic has roots, given as rationals by (1). Otherwise, there are no roots. So $2x^2-x-3=0$ does have two roots $x=\frac{1}{4}(1\pm5)=-1$ or $3/2$ by (1); but $x^2-\frac{1}{2}x-\frac{1}{4}=0$ has no roots since we cannot recognise the irrational values $x=\frac{1}{4}(1\pm\sqrt{5})$ given by (1). There is one odd case: if $b^2-4ac=0$, (1) gives $x=-b/2a$ and the equation has a rational root, but only one and not two. The difficulty is overcome by agreeing that there are two rational roots which happen to coincide when $b^2-4ac=0$.

We can do better if we allow the domain of x to be the set of all real numbers, rational or irrational. Then (1) gives two real roots provided that $b^2-4ac>0$, and two real roots which happen to coincide if $b^2-4ac=0$. The equation $x^2-\frac{1}{2}x-\frac{1}{4}=0$ is now accommodated, with roots $x=\frac{1}{4}(1\pm\sqrt{5})$. But we are still not able to say that every quadratic equation with real coefficients has two roots. The case where $b^2-4ac<0$ defeats us. For example, for the equation $x^2-\frac{1}{2}x+\frac{1}{4}=0$, (1) gives $x=\frac{1}{4}(1\pm\sqrt{-3})$ which we cannot recognise. We cannot wriggle out by saying that the roots are real if $b^2-4ac\geqslant0$ and 'imaginary' if $b^2-4ac<0$; we must know what 'imaginary' numbers are and extend the domain of x to include them. To justify the position, that every quadratic equation has two roots, the

* An interpretation is possible. The entry 0·30103 (to five decimal places) against $\log_{10}2$ can be interpreted: if $\log_{10}x=0\cdot30103$, then $x=10^{0\cdot30103}$ *exactly* and x is *close* to 2. It is the 2 which is an approximation; $\log_{10}2$ has no meaning but $\log_{10}x$ for certain x close to 2 does have meaning.

domain of x must be extended further and this requires more than elementary algebra.

(iv) Consider a pair of linear equations in two variables:

$$a_1x+b_1y+c_1=0 \quad \text{and} \quad a_2x+b_2y+c_2=0$$

where we take both real coefficients and domains of real numbers for the two variables x and y. Algebraic manipulation (Appendix A.4) gives:

$$x = \frac{b_1c_2-b_2c_1}{a_1b_2-a_2b_1} \quad \text{and} \quad y = \frac{c_1a_2-c_2a_1}{a_1b_2-a_2b_1} \quad \text{..................(2)}$$

We seem to have a very convenient result: every pair of linear equations has a unique solution, the real values of x and y being given by (2). There is no difficulty, as with the quadratic equation of (iii), about 'imaginary' values. But there is a difficulty. If $a_1b_2-a_2b_1 \neq 0$, the algebra is correct and (2) always gives real values for x and y. But what if $a_1b_2-a_2b_1=0$? The results (2) are then without meaning, since each denominator is zero. In fact, the algebra leading to (2) has gone astray and we are left with (apparently) no solution.

A graphical approach helps here. If we plot the relation

$$a_1x+b_1y+c_1=0,$$

as in Fig. 1.6, we get a line L_1; similarly from $a_2x+b_2y+c_2=0$ we get another line L_2. The values of x and y given by the point P of intersection of the two lines satisfy both relations, i.e. provide the solution of the pair of equations. The case illustrated has $x+y-3=0$ for L_1 and $2x-3y+3=0$ for L_2. The point of intersection has co-ordinates $\bar{x}=1{\cdot}2$ and $\bar{y}=1{\cdot}8$, the solution of the two equations, as can be checked from (2). The cases of failure are now evident. The lines L_1 and L_2 meet in a single point as long as they are distinct and not parallel; this is so when $a_1b_2-a_2b_1 \neq 0$. If the lines are distinct and parallel, then there is no point of intersection and we can say that the two equations have no solution because they are inconsistent. If the lines coincide, then there are indefinitely many points of 'intersection' and we can say that the two equations have an indeterminate solution because they are identical.

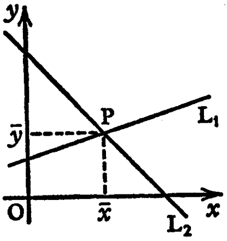


FIG. 1.6

Hence, as far as elementary algebra is concerned, the result is that the two linear equations have a unique solution given by (2), provided that $a_1b_2 - a_2b_1 \neq 0$. Cases where $a_1b_2 - a_2b_1 = 0$ cannot be handled. These are often called *degenerate cases*; they need to be examined further.

1.7. Notation. Mathematical exposition is greatly facilitated by a good notation. Conversely, a clumsy notation puts off the reader and tends to prevent the full exploitation of results. Notational difficulties are twofold. There is an acute shortage of letters and other symbols for use and a notation must be extremely economical. There is the need to ensure that a notation is simple, concise and pleasing — and at the same time generally accepted and understood. Not all notations are economical in the use of letters, nor are they always generally adopted; different writers may well use various notations for the same thing. However, a considerable degree of uniformity does exist and it is usually wise to stick to what is generally in use.

There are 26 letters in the Roman alphabet. The supply can be doubled by taking both small and capital letters and further increased by pressing the Greek alphabet into service:

Small	*Capital*	*Name*	*Equivalent*	*Small*	*Capital*	*Name*	*Equivalent*
α	A	alpha	a	ν	N	nu	n
β	B	beta	b	ξ	Ξ	xi	x
γ	Γ	gamma	g (hard)	o	O	omicron	o (short)
δ	Δ	delta	d	π	Π	pi	p
ϵ	E	epsilon	e (short)	ρ	P	rho	r
ζ	Z	zeta	z	σ	Σ	sigma	s
η	H	eta	e (long)	τ	T	tau	t
θ	Θ	theta	th	υ	Υ	upsilon	u
ι	I	iota	i	ϕ	Φ	phi	ph
κ	K	kappa	k	χ	X	chi	ch (hard)
λ	Λ	lambda	l	ψ	Ψ	psi	ps
μ	M	mu	m	ω	Ω	omega	o (long)

Certain letters or groups of letters are usually reserved for particular purposes. The later letters are used for variables: x, y, z and sometimes u, v, w; the Greek letters ξ, η, ζ are also employed for this purpose. For constants or parameters, it is usual to take early letters (a, b, c or the Greek α, β, γ) where the constant aspect is stressed; and to take middle letters (k, l, m, n or the Greek $\kappa, \lambda, \mu, \nu$) for a parametric interpretation. A few letters are kept almost entirely for particular purposes. The constants $\pi = 3{\cdot}14159 \ldots$ and $e = 2{\cdot}71828 \ldots$ are of such importance as almost to pre-empt these letters. Similarly Σ is reserved to denote summation. A natural notation for variable time is t or τ. In the calculus, d and D are used for derivatives, δ and $\varDelta$ for finite increments or differences. Almost invariably ϵ denotes a small constant, and θ often indicates a proper fraction (between 0 and 1). General expressions or functions lay claim to f, g, F, G and the Greek ϕ, ψ. These reservations, however, are subject to exceptions. For example, g may stand for the constant of gravity and F for a field (and not for a function), and θ can be used for an angle (and not a proper fraction).

These resources are still not enough. One way of stretching them further is to print in different types. Letters used as symbols are conveniently printed in italics, e.g. a and A, but different types are possible, e.g. **a** and **A** in bold. The difficulty here — and it prevents widespread use — is that what is possible in print is very difficult to convey in manuscript or typescript. Bold type can, however, be usefully introduced in such particular fields as matrix theory, and it is so used in Chapter 13 below.

Non-literal symbols are employed to economise on letters, mainly for relations between or operations on entities denoted by letters Well-known examples are $=$, $<$ and $>$ for 'equals', 'less than' and 'greater than' respectively. Variants on these are less familiar but very economical in use. So $\leqslant$ means 'less than or equal to' and $\geqslant$ 'greater than or equal to'. Further, $\neq$ means 'not equal to'; $\nless$ and $\ngtr$ have similar negative interpretations. Other non-literal symbols are involved in $n!$, $\binom{n}{r}$ and $|\, a \,|$ as defined below, and the symbol $\int$ appears in an integral. However, it is not helpful to clarity to scatter around large numbers of such symbols. Hence, some operations are denoted, not by non-literal symbols, but by an abbreviated form of the name of the operation, e.g. Lim for 'limit' and Max for 'maximum'.

There is one further extension possible, and it is a very important one: the use of numerical or literal subscripts. This provides a fine example of how the joint needs of economy and precision are satisfied. Much of mathematics deals with 'many' — either a specified but large number, or more usually an unspecified number, of things. Consider a set of constants, say the coefficients of a polynomial. There may, perhaps, be 4 of them, as in a cubic; they can be written a, b, c and d. There may be an unspecified number of them, as in a polynomial of nth degree, and they may be written $a, b, c, \dots k$. But this is both vague and wasteful. One letter, modified by subscripts, does much better: a_1, a_2, a_3, a_4 for four constants; $a_1, a_2, a_3 \dots a_n$ for n constants. The notation can then be condensed further, by using general subscripts (i and j, or r and s are commonly adopted) and by indicating the values they take. So, for 4 and n constants:

$$a_r \quad r=1, 2, 3, 4 \quad ; \quad a_s \quad s=1, 2, 3, \dots n.$$

These are simple sequences. A further development suggests itself, to accommodate double (or higher) sequences. A double array of $m \times n$ constants can be denoted by a single letter and two subscripts:

$$\begin{array}{ccccc} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ & \dots & \dots & \dots & \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{array}$$

where the first subscript indicates the row and the second the column. The array can then be drastically and conveniently condensed to

$$a_{rs} \quad r=1, 2, 3, \dots m \quad \text{and} \quad s=1, 2, 3, \dots n.$$

Triple arrays can be handled similarly, with three subscripts, and so on.

The use of subscripts makes it possible to denote sums of items in a very compact form, by means of the $\sum$ *notation*. Here the Greek capital $\sum$ stands for 'sum' the items which are indicated by the symbol or symbols following $\sum$. So, as a matter of notation and in the interests of brevity, write:

$$\sum_{r=1}^{4} a_r = a_1 + a_2 + a_3 + a_4$$

$$\sum_{r=1}^{n} a_r = a_1 + a_2 + a_3 + \dots + a_n.$$

The flexibility of the notation is seen in a few illustrations:

$$\sum_{r=1}^{3} a_r b_r = a_1 b_1 + a_2 b_2 + a_3 b_3$$

$$\sum_{r=1}^{n} a_r x_r^2 = a_1 x_1^2 + a_2 x_2^2 + a_3 x_3^2 + \dots + a_n x_n^2$$

$$y_1 = \sum_{s=1}^{n} a_{1s}x_1x_s = a_{11}x_1^2 + a_{12}x_1x_2 + a_{13}x_1x_3 + \ldots + a_{1n}x_1x_n.$$

Notice that the subscript r used in a sum like $\sum_{r=1}^{n} a_rb_r$ is 'knocked out' by the $\sum$; the expression obtained does not depend on r. In fact, it can be changed without altering anything; it is just a matter of a convenient label:

$$\sum_{r=1}^{n} a_rb_r = \sum_{s=1}^{n} a_sb_s = \sum_{t=1}^{n} a_tb_t = \ldots$$

are all the same thing $(a_1b_1 + a_2b_2 + a_3b_3 + \ldots + a_nb_n)$. Such a subscript is called a *dummy subscript*. On the other hand, there may be a *free subscript* which appears in every item and which is not subjected to a summation process. In the last illustration above, $y_1 = \sum_{s=1}^{n} a_{1s}x_1x_s$, the subscript 1 is a free one. If it is changed, say to 2, then a different expression $y_2 = \sum_{s=1}^{n} a_{2s}x_2x_s$ results. This suggests a further development. Take a sequence of free subscripts and write:

$$y_1 = \sum_{s=1}^{n} a_{1s}x_1x_s = a_{11}x_1^2 + a_{12}x_1x_2 + a_{13}x_1x_3 + \ldots + a_{1n}x_1x_n$$

$$y_2 = \sum_{s=1}^{n} a_{2s}x_2x_s = a_{21}x_1x_2 + a_{22}x_2^2 + a_{23}x_2x_3 + \ldots + a_{2n}x_2x_n$$

....

$$y_m = \sum_{s=1}^{n} a_{ms}x_mx_s = a_{m1}x_1x_m + a_{m2}x_2x_m + a_{m3}x_3x_m + \ldots + a_{mn}x_mx_n.$$

Then the general expression is:

$$y_r = \sum_{s=1}^{n} a_{rs}x_rx_s \quad r = 1, 2, 3, \ldots m$$

where r is the free subscript, as opposed to the dummy s. Finally, r can be 'knocked out', or made a dummy, by summing a second time:

$$\sum_{r=1}^{m} y_r = \sum_{r=1}^{m} \sum_{s=1}^{n} a_{rs}x_rx_s$$

which, on writing out in full, is a double array of terms, all added.

Caution is needed; the Σ notation is very concise and convenient but, until experience is gained in its use, it is easy to make errors. The summation process of Σ can deal easily with sums and constant factors:

$$\sum_{r=1}^{n}(a_r+b_r)=\sum_{r=1}^{n}a_r+\sum_{r=1}^{n}b_r\ ;\ \sum_{r=1}^{n}ka_r=k\sum_{r=1}^{n}a_r\quad (k \text{ constant}).$$

For example, the following are perfectly in order:

$$\sum_{r=1}^{n}(a_r+b_r)^2=\sum_{r=1}^{n}(a_r^2+2a_rb_r+b_r^2)=\sum_{r=1}^{n}a_r^2+2\sum_{r=1}^{n}a_rb_r+\sum_{r=1}^{n}b_r^2$$

and
$$\sum_{s=1}^{n}a_{rs}x_rx_s=x_r\sum_{s=1}^{n}a_{rs}x_s.$$

But care must be taken not to try to do this kind of thing for sums of products. $\sum_{r=1}^{n}a_rb_r=a_1b_1+a_2b_2+a_3b_3+\ldots+a_nb_n$ is not reducible in any way. In particular:

$$\sum_{r=1}^{n}a_rb_r\neq\left(\sum_{r=1}^{n}a_r\right)\left(\sum_{r=1}^{n}b_r\right)\ ;\ \sum_{r=1}^{n}x_r^2\neq\left(\sum_{r=1}^{n}x_r\right)^2$$

and
$$\sum_{r=1}^{n}(a_r+b_r)^2\neq\left(\sum_{r=1}^{n}a_r+\sum_{r=1}^{n}b_r\right)^2.$$

The operations of mathematics are denoted by symbols, the operators which say 'carry out the operation' concerned. It is usually the convention that, if several operations are made, the operators are to be read *from right to left*. The use of brackets, which are eliminated 'from the inside out' in simplification of an expression, helps considerably in keeping the order of operations. But brackets are often omitted, for brevity, and care is needed in (mentally) putting them back. For example, the operations of multiplication ×, of squaring $(\ldots)^2$ and of square root extraction $\surd(\ldots)$ are combined in:

$$\sqrt{xy^2}=\surd\{x\times(y)^2\}\quad(\text{reducing to } y\surd x \text{ when } y\geqslant 0).$$

This means: square y, multiply the result by x and take the root of the result. The order here cannot be changed; it is read from right to left as shown. It is quite a different thing to write:

$$x\surd(y^2)\quad(\text{reducing to } xy \text{ when } y\geqslant 0).$$

This becomes very important as more operators are added to the mathematician's armoury. In the calculus, the operator D means 'take the derivative of'. Consider a combination of D with the operator $\surd$ for 'take the square root of':

$$D\sqrt{1+x^2}=D\{\surd(1+x^2)\}$$

which means take the square root of $1+x^2$ and then take the derivative of the result. Calculus provides the answer: $\dfrac{x}{\surd(1+x^2)}$. This expression is *not* the same as:

$$\sqrt{D(1+x^2)}=\surd\{D(1+x^2)\}$$

which means take the derivative of $1+x^2$ and then take the square root of the result. The answer now is: $\sqrt{2x}$.

In conclusion, we can take note here of two special notations which are of frequent use in a variety of connections. The first provides a way of writing the product of all positive integers from 1 to n:

NOTATION: *The product of the integers from* 1 *to n is called* **n factorial**:

$$n!=n(n-1)(n-2)\ \dots\ 2\ .\ 1.$$

There are some obvious properties:

$$n!=n(n-1)! \quad \text{or} \quad \frac{n!}{(n-1)!}=n.$$

More generally, if r is a positive integer less than n:

$$n!=n(n-1)\ \dots\ (n-r+1)(n-r)! \ \text{ or } \ \frac{n!}{(n-r)!}=n(n-1)\ \dots\ (n-r+1).$$

Out of this there develops a further conventional notation:

NOTATION: $$\binom{n}{r}=\frac{n!}{r!(n-r)!}=\frac{n(n-1)\ \dots\ (n-r+1)}{r!}$$

where n and r are integers $(r<n)$.

In addition, it is sometimes convenient to write $\binom{n}{0}=\binom{n}{n}=1$. The expression $\binom{n}{r}$ is known in elementary algebra as the 'binomial coefficient' or as the 'number of combinations of n things r at a time', alternatively written ${}^nC_r$.

The other notation is for the magnitude of a real number, sign ignored:*

NOTATION: *The* **absolute value** *or* **modulus** *of a real number a is:*

$$|\,a\,|=\textit{positive number of pair }(a,\ -a)=\surd a^2.$$

It is immaterial whether $a=0$ is considered as a possibility here,

* The notation extends to the absolute value or modulus of a complex number (2.5 below).

since $| a | = 0$ can then be written if we wish. The following properties are derived:

$$| a | = a \ (a > 0) \quad \text{and} \quad | a | = -a \ (a < 0)$$
$$| ab | = | a | \times | b |$$
$$| a + b | \leqslant | a | + | b |.$$

Only the last causes any trouble; it is to be established by considering in turn all four cases obtained by taking a positive and negative and b positive and negative.

1.8. References. There are several short introductions designed to tell a general reader what mathematics is about and how its techniques are applied.

W. W. Sawyer: *Mathematician's Delight* (Pelican Books, London, 1943)

A. N. Whitehead: *An Introduction to Mathematics* (Home University Library, London, 1911)

are both successful in their rather different ways. As a first approach to the basic concepts of 'modern' mathematics, and therefore as a good preliminary reading before embarking on the present text, the following can be recommended:

Irving Adler: *The New Mathematics* (John Day, N.Y., 1958)

W. W. Sawyer: *Prelude to Mathematics* (Pelican Books, London, 1955); and *A Concrete Approach to Abstract Algebra* (W. H. Freeman, San Francisco, 1959).

The traditional courses on elementary mathematics in schools are reviewed at some length and re-interpreted in terms of basic ideas in:

W. L. Schaaf: *Basic Concepts of Elementary Mathematics* (John Wiley, N.Y., 1960)

The present volume is intended to take the reader farther, and more deeply, into basic mathematics than any of these introductory books do. It leaves plenty of scope for parallel reading and for exercises in various fields. The following represents a selection of the more specialist, but not very advanced, texts which can be used for the purpose. In algebra and finite mathematics:

J. G. Kemeny, J. L. Snell and G. L. Thompson: *Finite Mathematics* (Prentice-Hall, Englewood Cliffs, N. J., 1957)

D. C. Murdoch: *Linear Algebra for Undergraduates* (John Wiley, N.Y., 1957)

G. Birkhoff and S. MacLane: *A Survey of Modern Algebra* (Macmillan, N.Y., Revised Edition, 1953)

These texts have nothing to say about calculus, or about mathematical analysis generally. For this, good references are:

S. I. Altwerger: *Modern Mathematics* (Macmillan, N.Y., 1958)

R. Courant and H. Robbins: *What is Mathematics?* (Oxford University Press, 1941)

together with the ever-green and indispensible:

G. H. Hardy: *Pure Mathematics* (Cambridge University Press, 1st Edition, 1908; 10th Edition, 1952).

On linear systems, there is a stimulating if rather more advanced text:

R. A. Frazer, W. J. Duncan and A. R. Collar: *Elementary Matrices and some Applications to Dynamics and Differential Equations* (Cambridge University Press, 1947).

As the development proceeds, in the following chapters, brief references are given to the great mathematicians of the past and to the years in which they lived. These are intended as sign-posts for those interested in the historical evolution of the subject, an interest much to be encouraged. Parallel reading can be profitably undertaken in the history of mathematics, for example from:

E. T. Bell: *The Development of Mathematics* (McGraw-Hill, N.Y., 2nd Edition, 1945).

1.9. Exercises

1. The temperature of water is taken (at sea level) and the statement made: '$y°$ F is the same temperature as $x°$ C'. Express y in terms of x and specify the domain of x. (0° C corresponds to 32° F and 100° C to 212° F.)

2. Extract the four aces from a pack of playing cards and select x aces from the set of four. Let y be the number of different selections of x aces which can be made. Show that x has the domain $\{1, 2, 3, 4\}$ and y the range $\{1, 4, 6\}$

3. Draw a graph to show the expressions $y=x^2-2$ (x rational) and $y=4/x$ (x positive rational). Hence find one root of $x^3-2x-4=0$.

4. Plot $y=\sqrt{(4-x^2)}$ as a graph for real x, non-negative and not greater than 2 $(0 \leqslant x \leqslant 2)$. Find the solution sets for $\sqrt{(4-x^2)}=1$ and for $\sqrt{(4-x^2)}<1$.

5. Show that $y=|x|$ and $y=\sqrt{x^2}$ are identical expressions; represent graphically for real x, $-2 \leqslant x \leqslant 2$.

6. *nth Roots*. If a is a positive real number, the notation $\sqrt{a}$ stands for the *positive* square root of a. Why is no such qualification needed for $\sqrt[3]{a}$? What of $\sqrt[4]{a}$, $\sqrt[5]{a}$, ...?

7. Complete the square in $y=ax^2+bx+c$ and show that, if a and c have the

same sign and if $b^2 < 4ac$, then $y > 0$ for all real x (a and c positive) or $y < 0$ for all real x (a and c negative).

8. Since the solution of an equation is unchanged by the removal of a multiplicative factor, show that polynomial equations with integral and with rational coefficients are interchangeable concepts. Illustrate by showing that $8x^2 - 6x + 1 = 0$ and $x^2 - \frac{3}{4}x + \frac{1}{8} = 0$ are the same, with solution set$\{1/4, 1/2\}$.

9. Show that the quadratic $(4\pi - 1)x^2 - 6x - 8 = 0$ with real coefficients has two real roots $x = \{3 \pm \surd(32\pi + 1)\}/(4\pi - 1)$. Show that it arises in finding the radius (x feet) of a sphere with surface area equal to the area of a rectangle $(x + 2)$ feet by $(x + 4)$ feet. Establish that there is such a sphere, with radius a little over 1 foot.

10. *Solution of Cubic Equations*. Draw the graph of $y = 2x^3 - 3x^2 + 2$, insert $y = 1$, $y = \frac{3}{2}$, $y = 2$ and $y = 3$, and indicate why $2x^3 - 3x^2 - 1 = 0$ has only one and $2x^3 - 3x^2 + \frac{1}{2} = 0$ has three real roots. What of $2x^3 - 3x^2 + 1 = 0$ and $2x^3 - 3x^2 = 0$?

11. Show that $x^4 - x^2 - 2x + 2 = (x^2 - 1)^2 + (x - 1)^2$ is positive for all real x except that it has a zero when $x = 1$. Deduce that $x^4 - x^2 - 2x + 2 = 0$ has only two real roots, i.e. a double root $x = 1$.

12. If $x + y - 4 = 0$ and $x - y + 2 = 0$, show that $x = 1$ and $y = 3$. If $x + y - 4 = 0$ and $x - y + 2 > 0$, show that $y = 4 - x$ for $x > 1$. Illustrate graphically.

13. If k is a positive parameter, express the solution of $y = 1 - x$ and $y = x + k$ in terms of k. Restrict to $x \geqslant 0$, $y \geqslant 0$, and show that there is no solution unless $k \leqslant 1$. Interpret $y = 1 - x$ as the demand for a commodity at price x and $y = x + k$ as the supply, shifting with k. Illustrate graphically.

14. Show that $\sum\limits_{r=1}^{n} a_r x_r^2 > 0$ for all real $x_1, x_2, \ldots x_n$ if and only if $a_r > 0$ (all r).

15. From $\sum\limits_{r=1}^{n} a_r a_s = a_s \sum\limits_{r=1}^{n} a_r$, deduce $\sum\limits_{s=1}^{n} \sum\limits_{r=1}^{n} a_r a_s = \sum\limits_{s=1}^{n} \left(a_s \sum\limits_{r=1}^{n} a_r \right) = \left(\sum\limits_{r=1}^{n} a_r \right) \left(\sum\limits_{s=1}^{n} a_s \right)$

The last expression can be written $(\sum\limits_{r=1}^{n} a_r)^2$; why? If $n = 2$, show that all these double sums are simply: $a_1^2 + 2a_1a_2 + a_2^2$.

*16. Show that
$$\left(\sum_{r=1}^{n} a_r + \sum_{r=1}^{n} b_r \right)^2 = \left(\sum_{r=1}^{n} a_2 + 2 \sum_{r=1}^{n} a_r b_r + \sum_{r=1}^{n} b_r^2 \right) + A$$
where $A = \Sigma\Sigma a_r a_s + 2\Sigma\Sigma a_r b_s + \Sigma\Sigma b_r b_s$ $\left(\Sigma\Sigma = \sum\limits_{r=1}^{n} \sum\limits_{s=1}^{n} \text{ excluding } r = s\right)$.

Deduce that the difference $\left(\sum\limits_{r=1}^{n} a_r + \sum\limits_{r=1}^{n} b_r \right)^2 - \sum\limits_{r=1}^{n} (a_r + b_r)^2$ is A.

17. Illustrate the use of brackets by writing a^{b^c} *either* as $(a)^{b^c}$ *or* as $(a^b)^c = a^{bc}$. (The first is the usual interpretation.) As an instance, show that $2^{3^2} = (2)^{3^2} = 512$ and that $2^{3^2} = (2^3)^2 = 64$.

18. *Binomial Theorem*. Show that $(1 + x)^2 = 1 + 2x + x^2 = \sum\limits_{r=0}^{2} \binom{2}{r} x^r$ and that $(1 + x)^3 = 1 + 3x + 3x^2 + x^3 = \sum\limits_{r=0}^{3} \binom{3}{r} x^r$. Generalise to: $(1 + x)^n = \sum\limits_{r=0}^{n} \binom{n}{r} x^r$.

CHAPTER 2

NUMBER SYSTEMS

2.1. Rational numbers. A strictly logical treatment of mathematics starts with the general concept of sets on which all mathematics, and indeed formal logic, are based. It then proceeds, in an inevitably leisurely way, to groups and rings, to relations, mappings and transformations, all on an abstract level. It is some time before anything at all recognisable is reached; an attempt to drag in familiar constructions as illustrations could be made but it would be artificial and unconvincing. If interest is to be aroused and maintained, a compromise has to be reached between the desire for logical development from (unfamiliar) basic ideas on the one hand, and the need to keep in touch with (familiar) practical mathematics on the other hand. The present chapter and the next one represent such a compromise.

Much of mathematics, though by no means all, makes use of numbers. It is concerned with relations between variables taking numerical values. In application, it refers to 'objects' and 'entities' which are ordered and/or measured. Much of mathematics, but again not all, boils down in the end to solving equations, to finding the particular variables which satisfy given conditions: when will a ball thrown into the air start to come down, how much wheat will be sold at such a price, what is the particular path followed by the ball or by the price of wheat over time? With this in mind, we start here with the familiar number system, the numbers described as 'rationals'. At the same time, we bring in the equally familiar polynomial equations, specifically the linear, quadratic and cubic equations.

The treatment in these two chapters involves a good deal of jobbing backwards and forwards. Inevitably it will require tidying up later on when a more steady progress is made from the basic ideas of sets. There are, however, some very clear advantages. We are able to see more clearly, not only the correct ways of doing things, but also the reasons why they have to be done. Jobbing backwards shows up the need to take nothing for granted, to query everything, to delve deeper until the barest minimum of the undefined is exposed. Jobbing forwards avoids the awkward and unnecessary 'compartment' method of treatment. There is no need to exhaust one subject before tackling another; indeed there is much to be lost since mathematical techniques are highly inter-related. Further, when we do start on sets, groups and other basic concepts, we have plenty of material ready to hand for illustration.

Rational numbers comprise all the simple numbers dealt with in elementary arithmetic: positive and negative integers and fractions (ratios of integers), together with zero as the number separating the positive from the negative. The whole system of rationals can be looked upon either as a collection or *set*, or as an ordered *sequence*. (The concepts of sets and sequences here are the obvious ones; more specific developments of them come later.) The numbers are subject to the operations of addition and subtraction, of multiplication and division,* which we take for the moment as self-evident. As a set, the remarkable feature of the rationals is that they are *closed* with respect to the operations. Add or subtract any two rationals, multiply or divide them, and we still get a rational. The combination of rationals by means of $+$, $-$, $\times$ and $\div$ always gives another rational and never anything outside the set. As a sequence, the rationals are indefinitely extended, stretching in both directions through positive and negative numbers from 0. Further, between any two rationals (e.g. 0 and 1), there are as many others as we like to specify. They are uniquely ordered by greater and less. We can arrange any subset of them in ascending order, e.g.:

$$-\tfrac{3}{2},\ -1,\ -\tfrac{3}{4},\ -\tfrac{1}{8},\ 0,\ \tfrac{1}{3},\ \tfrac{1}{2},\ \tfrac{7}{8},\ 1,\ \tfrac{5}{4}.$$

It is easy to determine which of two rationals is the larger; for example $14/27 > 71/139$ since $14 \times 139 = 1946$ is greater than $27 \times 71 = 1917$.

There are some difficulties. As in nearly all number systems, the number zero needs careful handling, particularly as regards the operation of division. The way out here is explicitly to exclude the number zero from the multiplication/division process, keeping it for

* Provided that we do not divide by zero.

addition/subtraction. Another point arises in attempting to design a suitable notation. Rationals may be represented by p/q, where p and q are positive or negative integers, together with $p=0$. Whether this is satisfactory or not depends on the view taken about the ordering of rationals. If we are happy to have partial ordering according to greater, less and equals, with many rationals left on an equal footing, then the notation p/q serves. If we want, as is implicitly assumed above, a strict ordering according to greater and less, with no two rationals ranked equal, then it won't do. The notation shows

$$\frac{-2}{3}, \frac{2}{-3}, \frac{-4}{6}, \frac{4}{-6}, \frac{-6}{9}, \frac{6}{-9}, \ldots$$

as all different. They are equal from the ordering point of view; each can be represented by the single rational $-\frac{2}{3}$. There is the need to eliminate duplication and perhaps the simplest notation for what we have in mind is: $\pm p/q$, where p and q are relatively prime† positive integers together (as before) with $p=0$.

This discussion illustrates that, though arithmetic can proceed with fractions, it tends to be messy. An alternative is to work with the decimal notation, generally simpler but with a different kind of complication — the need to handle recurring decimals (2.9 Ex. 2).

2.2. The operational rules and order properties. The set of rational numbers is denoted by R. There are two operations, addition $(+)$ and multiplication $(\times)$, each applied to two members of R to give another member of R. The processes of subtraction and division are subsidiary (see 2.9 Ex. 3) to be derived from addition and multiplication respectively. The rules of addition and multiplication are so well-known that they are applied, even in the most elementary arithmetic, without thought. However, they had to be learnt, along with the addition and multiplication tables (which define the operations), at some time in everyone's early life. It is a very salutary exercise, and a very useful one in the subsequent development, to specify them precisely and to label them carefully for recognition later on. They are:

† 'Relatively prime' means no common factor other than 1, see 3.1.

The Operational Rules of Algebra
For the set $R=\{a, b, c, \ldots\}$ of rational numbers

Rule	*Addition* (+)	*Multiplication* (×)
1. Closure	$a+b$ belongs to R	$a \times b$ belongs to R
2. Associative	$a+(b+c)=(a+b)+c$	$a \times (b \times c)=(a \times b) \times c$
3. Commutative	$a+b=b+a$	$a \times b=b \times a$
4. Identity	Zero 0 such that $a+0=0+a=a$	Unity 1 such that $a \times 1=1 \times a=a$
5. Inverse	Negative $(-a)$ such that $a+(-a)=(-a)+a=0$	Reciprocal a^{-1} such that $a \times a^{-1}=a^{-1} \times a=1$ $(a \neq 0)$
5*A*. Cancellation	If $a+b=a+c$ then $b=c$	If $a \times b=a \times c$ $(a \neq 0)$ then $b=c$
6. Distributive	$a \times (b+c)=a \times b+a \times c$ and $(a+b) \times c=a \times c+b \times c$	

Here a, b and c are any rational numbers, belonging to R. Rule 1 (closure) expresses the fact that R is self-contained, both for + and for ×. Rule 2 (associative) indicates that the order of repeated addition or multiplication is immaterial. Rule 3 (commutative) indicates that, when two rationals are combined (by + or by ×), it is a symmetrical operation: a with b is the same as b with a. This reflects the fact that the operations are defined in a symmetric way. Rule 4 (identity) shows that R contains a unique member which doesn't alter another member by addition, and another unique member for multiplication. Rule 6 (distributive) links addition and multiplication.

Rule 5 (inverse) is the one which allows subtraction (or division) to be introduced, as undoing what is done by addition (or multiplication). Every rational has its negative; the two add to zero. Now *define*:

$$a-b=a+(-b)$$

where $(-b)$ is the negative of b and the algebraic process known as subtraction is accommodated. Similarly, *define* division in terms of reciprocals:

$$a/b=a \times b^{-1}$$

where b^{-1} is the reciprocal of b. The rule labelled 5*A* (cancellation) is added in the table for particular reasons. *First*, it is a very important consequence of 5 (together with the previous rules*). For multiplication:

Given: a has reciprocal a^{-1} $(a \neq 0)$.

From: $a \times b = a \times c \quad (a \neq 0)$

we get: $a^{-1} \times (a \times b) = a^{-1} \times (a \times c)$

i.e. by 2: $(a^{-1} \times a) \times b = (a^{-1} \times a) \times c$

i.e. by 5: $1 \times b = 1 \times c$

i.e. by 4: $b = c$ which is 5*A*.

Notice that it is in the writing of a reciprocal (rule 5), and equally in cancelling a common element (rule 5*A*), that the exception $a \neq 0$ must be made. Another version of the cancellation rule is obtained from 5*A* by putting $c = 0$:

5*B*. If $a \times b = 0$, then *either* $a = 0$ *or* $b = 0$.

Second, rule 5 is one which, for other systems, may very well not hold. It is then important to know whether 5*A* holds or not. Though 5*A* follows from 5, the converse is not so: if 5*A* holds, 5 need not. Hence 5*A* is a weaker version of 5. We may still be able to cancel, even if we cannot write inverses. On the other hand, neither may hold in which case we have the (apparently curious) situation, the negative of 5*B*:

There are non-zero a and b such that $a \times b = 0$.

This is described by saying that there are 'divisors of zero': 0 divided by b is a, and 0 divided by a is b.

Fortunately, the set R of rationals is well-behaved; it has reciprocals, cancellation can be done, and there are no divisors of zero. Such a set of numbers, for which the whole list of operational rules holds, is called a *field*. The field of rational numbers is the first of several fields we shall meet.

Rational numbers have the equally important property of being ordered; the set R is not only a field, it is an *ordered field*. The properties of order are well-known in elementary arithmetic but it is again a useful exercise to spell them out and to label them. The order symbol '$<$' is used for 'less than':

* Only the commutative rule 3 is not used, an important point later.

The Properties of Order

For the set $R=\{a, b, c, \ldots\}$ of rational numbers

Name	*Property*
(i) Trichotomy	One and only one of $a<b$, $a=b$, $b<a$ holds
(ii) Transitivity	If $a<b$ and $b<c$, then $a<c$
(iii) Density	If $a<b$, then c exists so that $a<c<b$
(iv) Extension	For any a, b exists so that $a<b$ and c exists so that $c<a$
(v) Consistency	If $a<b$, then $a+c<b+c$ for any c and $a\times c<b\times c$ for any positive c

There is an alternative symbol: '$>$' for 'greater than'. It is in common use but it adds nothing to the set of properties. Switch a and b around and '$<$' is replaced by '$>$': if $a<b$ then $b>a$. The transitive property, for example, then reads: if $a>b$ and $b>c$, then $a>c$.

The development of real numbers, in 2.4 below, is based on the concept of the rationals as an ordered sequence. The properties of order are further examined, in more general contexts, in connection with ordered fields (6.7 below) and in specifying order as a relation (7.7. below).

2.3. A wider field of numbers. Rational numbers serve for most purposes of arithmetic but are soon found to be inadequate in algebra, e.g. in solving quadratic equations. Consider the 'number' $\surd 2$, the extraction of the square root of 2 in the solution of the simple quadratic $x^2-2=0$. This 'number' arises in many ways; for example, by Pythagoras' Theorem, it is the length of the diagonal of a square of unit side. It can be shown, quite formally, that $\surd 2$ is *not* a rational number, as in the following *reductio ad absurdum* proof. Suppose $\surd 2$ is rational and write it p/q for certain integers p and q. Then $(p/q)^2=2$, or $p^2=2q^2$. If q is odd, q^2 has no factor 2; if q is even, q^2 has an even number of factors 2. Hence, $p^2=2q^2$ has 2 as a factor an odd number (1, 3, 5, ...) of times. This is impossible, whether p is odd or even. Hence $\surd 2$ is not rational.

Many other polynomial equations fail to have rational solutions in the same way as $x^2-2=0$. This is an unsatisfactory situation, needing correction. The way out, in elementary arithmetic, is to carry out certain computations and to write, for example, $\surd 2=1{\cdot}414$.

But this is merely a rational approximation $1414/1000 = 707/500$ to $\surd 2$; it is *not* $\surd 2$ itself. What is needed, clearly, is an extension of the system of rational numbers, so that equations like $x^2 - 2 = 0$ can be solved, and so that the operational rules and properties of 2.2 are preserved. The full extension is given in 2.4; meanwhile here is a method which suggests itself.

The root of $x^2 - 2 = 0$ is not a rational in the field R. *Define* it as a number of a new kind and write it $\surd 2$. (Strictly, there are two roots $\pm \surd 2$, but it is enough to take the positive $\surd 2$ and allow the negative $-\surd 2$ to arise automatically.) To all the rationals of R, throw in the new number $\surd 2$ and form all the necessary sums and products to ensure that the wider set of numbers is still a field, satisfying the rules of 2.2. This process, if it can be carried through, is called the *adjunction* of the new element $\surd 2$ to the field R. It produces a new and wider field, denoted $R(\surd 2)$. It is a very general and useful procedure, and other examples are met later.

It remains to show that the object can be achieved in this case. The new number $\surd 2$ must combine with the old ones (rationals); we must consider $a + b\surd 2$ as a new number*, multiplying $\surd 2$ by the rational b and then adding the rational a. But this is not all. We still have to combine the new numbers amongst themselves, to ensure closure in the wider set. This requires a *re-definition* of addition and multiplication, for numbers of the form $a + b\surd 2$. The definition is not arbitrary, since it is designed with a sharp eye on what we want to achieve:

$$\left.\begin{aligned}(a+b\surd 2)+(c+d\surd 2)&=(a+c)+(b+d)\surd 2\\(a+b\surd 2)(c+d\surd 2)&=ac+(ad+bc)\surd 2+bd(\surd 2)^2\\&=(ac+2bd)+(ad+bc)\surd 2\end{aligned}\right\}\quad\ldots\ldots(1)$$

since by definition $(\surd 2)^2 = 2$. We have now achieved closure, for the sums and products defined by (1) are themselves of the same form:

$$\text{(rational number)} + \text{(rational number)} \times \surd 2.$$

Finally, we check all the other rules of 2.2, one by one, to see that they still hold for $R(\surd 2)$ made up of numbers of form $a + b\surd 2$. It is found that they do (2.9 Ex. 6). Take rules 4 and 5 for multiplication

* Here and later, if no ambiguity arises, we drop the × sign for multiplication. Hence, ab means $a \times b$, $a + b\surd 2$ means $a + b \times \surd 2$, and so on.

as an illustration. The unity of $R(\surd 2)$ is still 1 (i.e. $1+0\ \surd 2$) since putting $c=1$, $d=0$ in (1) gives:

$$(a+b\surd 2)1=a+b\surd 2.$$

To get the reciprocal of $(a+b\surd 2)$, notice that another application of (1) gives:

$$(a+b\surd 2)(a-b\surd 2)=a^2-b^2(\surd 2)^2=a^2-2b^2$$

or:

$$(a+b\surd 2)\,\frac{a-b\surd 2}{a^2-2b^2}=1.$$

Hence $a+b\surd 2$ has a reciprocal $\dfrac{a-b\surd 2}{a^2-2b^2}=\left(\dfrac{a}{a^2-2b^2}\right)+\left(\dfrac{-b}{a^2-2b^2}\right)\surd 2$, which is of the same form.

A new field of numbers $R(\surd 2)$, comprising all numbers of the form $a+b\surd 2$ (a and b rationals), is thus created. It contains the old field R of rationals (the special cases with $b=0$), it contains many new numbers (those involving $\surd 2$), it satisfies all the operational rules and it can be operated upon algebraically in the familiar way. Some quadratic equations can now be solved, with roots in $R(\surd 2)$. One of them is $x^2-2=0$, with roots $x=\pm\surd 2$. Another is

$$2x^2-4x+1=0.$$

The process of completing the square gives:

$$2x^2-4x+1=2(x^2-2x+1)-1=2(x-1)^2-1$$

and the equation becomes:

$$2(x-1)^2=1 \quad\text{or}\quad (x-1)^2=2/4 \quad\text{or}\quad x-1=\pm\tfrac{1}{2}\surd 2.$$

The roots are $x=1\pm\frac{1}{2}\surd 2$, of the form $a+b\surd 2$ of the field $R(\surd 2)$.

Unfortunately, the wider field $R(\surd 2)$ does not take us very far, even with quadratic equations. There are many still without solutions; $x^2-3=0$ is a simple case. We could attempt to pursue the line of development: the adjunction of $\surd 2$ to the field R gives a wider field of numbers $R(\surd 2)$; the adjunction of $\surd 3$ to $R(\surd 2)$ gives a still wider field of numbers $R(\surd 2, \surd 3)$; and so on. This turns out to be an impossibly protracted line. Having noted that a wider *field* can be got by adding a new number to an existing *field*, we look for some more powerful and corner-cutting method of extending the number system. The extension adopted (2.4) is a difficult one to achieve. The subsequent further extension (2.5) then gets by with the adjunction of a single new number on the lines here indicated.

2.4. Real numbers. View the set R of rational numbers as an ordered sequence with the order properties of 2.2. R is indefinitely 'dense' in the sense of property (iii): select two rationals a and b, no matter how close, and there is always room for a third rational c to fall between then, for example $c=\frac{1}{2}(a+b)$. Further, R is indefinitely 'extended' in the sense of property (iv): for any rational a, no matter how large, there is always a larger $b>a$ (and similarly the other way).

It is useful as an illustration to represent rational numbers as points on a line drawn horizontally as in Fig. 2.4a. This is a *directed line*, increasing rationals being shown by points arranged in order from left to right. Since R is indefinitely extended both ways, the line must run indefinitely to the left and to the right from O, the point selected to represent zero. The property of indefinite density corresponds to the fact that a point can be inserted on the line between any two given points. To illustrate, take mid-points and insert A_1 between A and B, then A_2 between A_1 and B, then A_3 between A_2 and B, ... The rationals of R are shown in order as points on the line, yet indefinitely extended both ways and indefinitely dense.

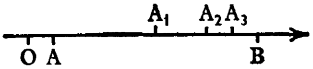


Fig. 2.4a

We need an extension to allow for 'numbers' like $\surd 2$ which are not rational. An arithmetic approach would be: represent a rational p/q as a decimal. If q has only 2's and 5's as factors (apart from those also factors of p), the decimal terminates, e.g. $71/50=1\cdot 42$. Otherwise the decimal recurs, e.g. $22/7=3\cdot\dot{1}4285\dot{7}$. It is very tempting to say that a decimal which neither terminates nor recurs gives a new number, an 'irrational', not represented in the form p/q. The process of root extraction gives $\surd 2=1\cdot 4142\ldots$ without terminating or recurring, so that $\surd 2$ is an 'irrational'. The whole set of rationals and 'irrationals' together then comprise the set of 'real numbers', wider than the set of rationals.

The difficulty is that we may never know for certain whether a decimal stops or recurs. The definition suggested is too loose to serve as a sound basis for real numbers; and we *do* need a very firm foundation, even for purposes of practical mathematics. It is not only a matter of tidying up the solution of quadratics and other equations. The decisive consideration is that real numbers are essential for the notion of a limit and so for breaking through from algebra to the calculus. Without a firm basis for real numbers, we cannot hope to establish the concept of a limit and, hence, to bring in the whole of the powerful apparatus of the calculus.

The strict definition of a real number is one of the more difficult exercises in basic mathematics. Consider three variants of what is essentially the same process. The first uses a *sequence of rational numbers* and the idea that a sequence may have a *limit*. It brings in, therefore, the idea of a limit, so important later on, and does so explicitly. A particular form of sequence suggests itself: the familiar process of successive approximation to a number by taking more decimal places. Two cases illustrate:

(i) The rational 22/7 can be approximated from below by the *increasing* sequence:

$$3{\cdot}1;\ 3{\cdot}12;\ 3{\cdot}14;\ 3{\cdot}141;\ 3{\cdot}142;\ \ldots$$

or from above by the *decreasing* sequence:

$$3{\cdot}2;\ 3{\cdot}17;\ 3{\cdot}15;\ 3{\cdot}145;\ 3{\cdot}143;\ \ldots$$

Represent these rationals, in pairs, on successive lines, as in Fig. 2.4*b*. The first pair is 3·1 and 3·2; the second pair is 3·12 and 3·17; and so on. Either sequence, continued indefinitely, tends to 22/7 in the limit. So does any 'crossed' sequence such as:

$$3{\cdot}1;\ 3{\cdot}17;\ 3{\cdot}14;\ 3{\cdot}145;\ 3{\cdot}142;\ \ldots$$

which is neither increasing nor decreasing.

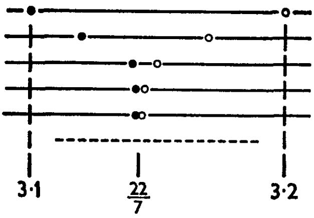


FIG. 2.4*b*

(ii) The irrational $\sqrt{2}$, i.e. the positive number a such that $a^2=2$, can be approximated by an increasing sequence from below:

a	1·4	1·40	1·41	1·412	1·414	...
a^2	1·96	1·96	1·9881	1·993744	1·999396	...

or by a decreasing sequence from above:

a	1·5	1·45	1·42	1·417	1·415	...
a^2	2·25	2·1025	2·0164	2·007889	2·002225	...

or by a 'crossed' sequence such as:

$$1{\cdot}5;\ 1{\cdot}40;\ 1{\cdot}42;\ 1{\cdot}412;\ 1{\cdot}415;\ \ldots$$

All sequences tend to the same limit, which represents $\sqrt{2}$. A diagram similar to Fig. 2.4*b* can be drawn in this case.

The suggestion here is that, if a sequence of rationals has a limit, then the limit may also be a rational; but it can equally well be an irrational. There is a *first possible definition* of a real number:

A *real number* α is the limit of a sequence of rational numbers a_n as the integer n increases indefinitely.

The system of real numbers, on such a definition, is wider than but includes all rationals. The additional cases are irrationals such as $\sqrt{2}$.

The second variant uses a *sequence of nested intervals* so arranged that they shrink to a *final residue* which is a single point. This is a slight but definite improvement on the first variant, as we see when we come to define a limit. Write the interval $[a, b]$ to denote all rationals x such that $a \leqslant x \leqslant b$. Take a sequence of nested intervals, each contained within the previous one. In case (i), such a sequence is:

$$[3{\cdot}1, 3{\cdot}2];\ [3{\cdot}12, 3{\cdot}17];\ [3{\cdot}14, 3{\cdot}15];\ [3{\cdot}141, 3{\cdot}145];\ [3{\cdot}142, 3{\cdot}143];\ \ldots$$

as represented in Fig. 2.4*c*. In case (ii) the sequence:

$$[1{\cdot}4, 1{\cdot}5];\ [1{\cdot}40, 1{\cdot}45];\ [1{\cdot}41, 1{\cdot}42];\ [1{\cdot}412, 1{\cdot}417];\ [1{\cdot}414, 1{\cdot}415];\ \ldots$$

would serve; it can be represented on a diagram similar to Fig. 2.4*c*. In each case, the sequence of nested intervals shrinks to a single point, which may be a rational (e.g. 22/7) or an irrational (e.g. $\sqrt{2}$). This suggests a *second possible definition* of a real number:

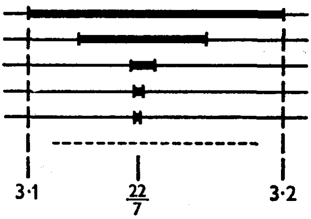


FIG. 2.4*c*

A real number α is the final residue of a sequence of nested intervals $[a_n, b_n]$ which shrink to a single point as n increases indefinitely; α is contained in all intervals.

Again the system of real numbers, so defined, extends and includes the rationals.

The third variant proceeds by specifying a *cut* (or section) of the set R of rationals into two parts, L and G, such that L is less than (to the left of) G. The cut can be made in any way provided that:

(1) each rational is in L or in G

(2) each of L and G contains at least one rational

and

(3) each rational of L is less than each rational of G.

An obvious way to specify a cut is by an inequality. In case (i), take L as all rationals x such that $7x < 22$ and G as all rationals x such that $7x \geqslant 22$. There is a dividing point between L and G, the point α *at* the cut, and here $\alpha = 22/7$. It happens that α belongs to G in this specification. This is accidental; a slightly different cut is got by taking x in L if $7x \leqslant 22$ and in G if $7x > 22$, and $\alpha = 22/7$ belongs to L. In case (ii), take all negative rationals in L together with all positive rationals x such that $x^2 < 2$; G comprises all positive rationals x such that $x^2 > 2$. There is again a dividing point α, but it is not a rational, neither in L nor in G. It is an irrational number, in fact $\surd 2$, which plugs a gap left in the rationals between L and G. This suggests a third possible definition of a real number, the one we adopt:

Definition: *A* **real number** α *is the point dividing the parts L and G of a cut of the set R of rationals satisfying* (1), (2) *and* (3) *above.* The system of real numbers so defined again includes the rationals.

The definition adopted has advantages over the others suggested. The cut specified includes and extends the idea of a sequence. As illustrated in Fig. 2.4*d*, an increasing sequence of rational L_1, L_2, L_3, L_4, ... approximating to α from below is simply a sequence selected from L; similarly for a decreasing sequence in G. Hence, L or G serves to fill out particular sequences, to comprehend them all. The reason for adopting the definition, however, is that it is best designed to establish the properties of real numbers:

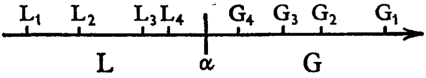


Fig. 2.4*d*

(*a*) Addition and multiplication can be re-defined for real numbers to satisfy the same operational rules as for rationals, i.e. real numbers form a *field*.

(*b*) Order by $<$ (less than) can be re-defined for real numbers to have the same order properties as for rationals, i.e. real numbers form an *ordered field*.

(*c*) If a cut is made in the set of real numbers, the dividing point is always a real number and nothing new is produced, i.e. the real numbers form a *complete ordered field*.

The formal definition of a real number, and an indication of how the properties (*a*), (*b*) and (*c*) are established, are given in 15.1.

We assume here that this is all achieved. We have then the complete ordered field of real numbers, denoted R^*. The approach we have adopted makes the properties appear sensible. Properties (*a*) and (*b*) imply that real numbers *extend* the system of rational numbers (filling the gaps in them) and that, at the same time, they *preserve* the operational rules and order properties of rationals. Property (*c*) is the new idea; it is the result of a basic theorem derived from the definition of real numbers. It is a property much to be desired. The implication is the following. If a cut of the *rationals* is made, a rational may be produced, but equally a new (irrational) number. Then, if a cut of the *real numbers* is made, nothing more emerges, just a real number. The real numbers are complete; no more gaps remain to be filled. We have reached the end of a line of development.

The property of completeness, distinguishing real from rational numbers, is the one required to achieve the passage from algebra to calculus. It is matched by a corresponding completeness of points on a directed line. We have the *continuum* of real numbers, of points on a line. The completeness property is to be put into another form, one of particular convenience. First, a definition:

DEFINITION: *A set S of real numbers has a* **lower bound** *a' if $a' < x$ for each x in S, and a* **greatest lower bound** *(GLB) a if no real number $x > a$ is a lower bound of S. An* **upper bound** *b' and a* **least upper bound** *(LUB) b are defined similarly.*

Sets of real numbers need not have a GLB or LUB. For example, the set of positive real numbers has no LUB and the set of integers has no GLB or LUB. But, if a set S of real numbers does have GLB a and LUB b, then $a \leqslant x \leqslant b$ for all x of S. Notice, however, that a and b themselves may or may not belong to S. If S has GLB a, then S may contain a (its least member), but it may not. For example, S consisting of x such that $x^2 > 2$ has GLB $a = \surd 2$, not contained in S. Vary slightly and specify S as comprising x such that $x^2 \geqslant 2$; S still has GLB $a = \surd 2$, now contained in S. It is of little significance whether a set includes its GLB (or LUB) or not; it is a matter which must not be allowed to cause any confusion.

The following result is a consequence of the complete ordered property of the set R^* of all real numbers:

THEOREM: *If a set S of real numbers has a lower bound, then it has a GLB; if S has an upper bound, then it has a LUB.*

This result should now appear both obvious and useful. To see what we have achieved, however, notice that a similar result is *not* true of rationals. The set of rationals x such that $x^2>2$ certainly has a lower bound (e.g. 0 or 1) but it has no *rational* as GLB; if a (rational) is *any* lower bound ($a^2 \not> 2$) there are always rationals $b>a$ such that $b^2 \not> 2$, i.e. b is also a lower bound. It takes the definition of the real number $\sqrt{2}$ to provide a GLB. See 2.9 Ex. 9.

What do we mean when we say that we have now taken a decisive step forward? Certainly, we have provided a proper conceptual basis for much of practical arithmetic. A square of unit side has diagonal $\sqrt{2}$; the area of a circle of radius r is πr^2. In dealing with these things, we cheat a little; we approximate $\sqrt{2}=1{\cdot}414$ and $\pi=22/7$, or whatever we find convenient. Moreover, in drawing the graph of the curve $y=x^2-2$, we again cheat by giving x a suitable sequence of rational values, by finding the corresponding rational values of y and finally by putting a *smooth* curve through the plotted points. We can now see the extent of the cheating involved. But the reason why the step forward is so decisive is that the way is clear for an advance into new territory, that of the calculus and mathematical analysis. It is essential to write y as a function of x, where x is a continuous or real variable; rational values of x will just not do. This is as essential in practice as in theory. Practical mathematics, as used by the physicist or engineer, cannot do without real numbers.

2.5. Complex numbers. The real numbers are a completely ordered set, shown by the continuum of points on a line. There is nothing to be added to the order of the numbers or points. If we go further, we must give up the order by greater and less and we must go outside the line of one dimension. It is easily seen, however, that we do need to go further. Some quadratic equations do have real roots, e.g. $2x^2-x-3=0$ gives $x=3/2$ or -1 and $x^2-2=0$ gives $x=\pm\sqrt{2}$. But we are still left with some quadratics on our hands, e.g. $x^2+x+1=0$, with roots which may be described as 'imaginary' (1.6 above). In this unsatisfactory situation, the obvious line to take is to extend the number system, again in such a way that the operational rules of 2.2 are preserved. The object is to extend the field R^* of real numbers to a wider field in which all equations are solved. The new field will not be ordered; it cannot be, since R^* is complete. But the price, the loss of order, is worth paying.

The neatest way of extending R^* is by use of the idea of adjunction (2.3). Since $x^2+1>0$ for all real x, the equation $x^2+1=0$ has no real solution. Introduce a new number i which is a root of the equation, so that $i^2=-1$. Into the set R^*, closed under the operations of addition (+) and multiplication (×), we throw the new number i. We impose the same operations of + and × on combinations of i with real numbers $a, b, c, d, \ldots$. We write $a+ib=a+i\times b$ and we add this to, or multiply it by, another of its kind $(c+id=c+i\times d)$ according to the rules:

$$\left.\begin{aligned}(a+ib)+(c+id)&=(a+c)+i(b+d)\\(a+ib)\times(c+id)&=ac+i(ad+bc)+i^2bd\\&=(ac-bd)+i(ad+bc)\end{aligned}\right\}\quad\ldots\ldots\ldots\ldots\ldots(1)$$

This amounts to a *re-definition* of sums and products to accommodate the new number i, and it makes use of the fact that i is the number such that $i^2=-1$.

Consider the set of numbers $a+ib$, where a and b are any real numbers and where $i^2=-1$. Addition and multiplication are closed for this set, since (1) shows that adding (or multiplying) two such numbers gives another number of the same form. The unity of the new set is still $1=1+i0$ since:

$$(a+ib)\times 1=a+ib \quad \text{by (1) with } c=1,\ d=0.$$

Further, another application of (1) gives: $(a+ib)\times(a-ib)=a^2+b^2$.

Hence: $$(a+ib)\left\{\frac{a}{a^2+b^2}+i\left(\frac{-b}{a^2+b^2}\right)\right\}=1$$

and the reciprocal of $a+ib$ exists as $\dfrac{a}{a^2+b^2}+i\left(\dfrac{-b}{a^2+b^2}\right)$, a member of the same set. We can write:

$$(a+ib)^{-1}=\frac{1}{a+ib}=\frac{a-ib}{(a+ib)(a-ib)}=\frac{a-ib}{a^2+b^2}$$

on clearing the denominator of i (see Appendix A.6). All the other operational rules of 2.2 apply, as can easily be checked (2.9 Ex. 6). The set of numbers $a+ib$ forms a field.

The object is achieved. The adjunction of just one new element i (where $i^2=-1$) to the field R^* of real numbers, with + and × redefined as in (1), produces a new field of numbers $a+ib$ (a and b real). We have the field C of *complex numbers* $a+ib$. C includes R^*, since $a+ib$ is the real number a when $b=0$.

The process of solving the quadratic equation $ax^2+bx+c=0$ is now complete. The solution offered in 1.6, $x=\{-b\pm\surd(b^2-4ac)\}/2a$, gives two real values when $b^2>4ac$ (coincident when $b^2=4ac$). It gives two complex values $x=\{-b\pm i\surd(4ac-b^2)\}/2a$ when $b^2<4ac$. For example, $x^2+x+1=0$ has roots: $x=(-1\pm i\surd3)/2$.

The field C has a remarkable property. The fundamental theorem of algebra (Chapter 3) shows that *all* polynomial equations† are solved within the field of complex numbers $a+ib$, including real numbers ($b=0$). Once again we are at the end of the line. For ordinary algebra, we require more than the field R^* of real numbers, but the field C of complex numbers is the biggest we need. The adjunction of just one new element, i as a root of $x^2+1=0$, turns the trick. We can then solve, not only quadratics, but *all* polynomial equations. On the other hand, though R^* is ordered, C is not. We cannot place i, the new number, in any order amongst real numbers. Graphically, i cannot be shown on the line of real numbers; if it is to appear at all, it requires an extra dimension.

The definition of complex numbers on these lines is neat but abstract. We get no idea of how complex numbers can be applied, or of their graphical representation — apart from the negative result that they need an extra dimension. Hence we give, as our definition of complex numbers, a more pedestrian concept, but one which enables us to bring in two constructions of the utmost importance in themselves: vectors and number pairs.

DEFINITION: *The* **complex number** z *is the pair* (x, y) *of real numbers subject to the rules for equality* (=), *sums* (+) *and products* (×):

$$\left.\begin{array}{r}(x_1, y_1)=(x_1, y_1) \quad \textit{implies} \quad x_1=x_2 \text{ and } y_1=y_2\\ (x_1, y_1)+(x_2, y_2)=(x_1+x_2, y_1+y_2) \qquad\qquad\qquad\\ x_1, y_1)\times(x_2, y_2)=(x_1x_2-y_1y_2, x_1y_2+x_2y_1) \qquad\end{array}\right\}\quad\ldots\ldots\ldots\ldots(2)$$

† With rational, or indeed real or complex, coefficients.

The rules (2) for sums and products are by no means arbitrary. On the contrary, they are cunningly devised to agree with (1), i.e. so that the operational rules of 2.2 are obeyed. Later, we have little difficulty in establishing that the complex numbers z, so defined, form a field.

A graphical interpretation is given in Fig. 2.5*a*, a representation known as an *Argand Diagram*, after Argand (1768–1822). The ordering of the number pair $z=(x, y)$ is essential: x first, y second. The pair then serves as the co-ordinates of a point P in a plane, referred to axes Oxy and with a scale of measurement fixed on each. The unit circle (centre O and unit radius) is drawn to indicate magnitudes; it cuts the axes in A, B, A' and B'. The complex number z is shown either by the *point* P (x, y) or by the *vector* OP. P is located by taking $OM=x$ and $ON=y$. Alternatively, the location of P is by the length r of the vector OP and the angle θ it makes with Ox. Here θ is a number, the measure of the angle in convenient units (e.g. in degrees). By elementary trigonometry, $x=r\cos\theta$ and $y=r\sin\theta$ so that $r^2=x^2+y^2$ and $\tan\theta=y/x$ (see Appendix A.9). Hence:

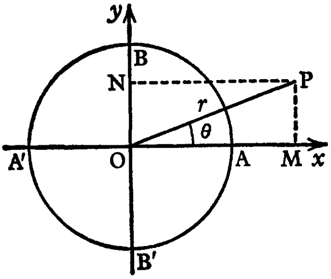


FIG. 2.5*a*

NOTATION: *The* **absolute value** *or modulus of the complex number* $z=(x, y)$ *is* $r=\sqrt{(x^2+y^2)}$; *and its* **argument** *or amplitude is the angle* θ *such that* $x=r\cos\theta$ *and* $y=r\sin\theta$.

The absolute value is also written $|z|$, as indicated in 1.7 above.

The rules (2) imposed on complex numbers have their graphical interpretation on an Argand Diagram (see 2.9 Ex. 11 and 13). First, it takes *two* things to fix the location of a point in a plane. Hence, $z_1=(x_1, y_1)$ is equal to $z_2=(x_2, y_2)$ only if *both* conditions, $x_1=x_2$ and $y_1=y_2$, hold; the corresponding points P_1 and P_2 coincide only if both co-ordinates are the same. Second, the vectors OP_1 and OP_2 add to the vector OP (for $z=z_1+z_2$), where OP is the resultant of OP_1 and OP_2 in the sense (familiar in mechanics) that OP is the diagonal of the parallelogram formed from OP_1 and OP_2 (Fig. 2.5*b*). Third,

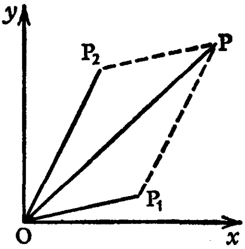


FIG. 2.5*b*

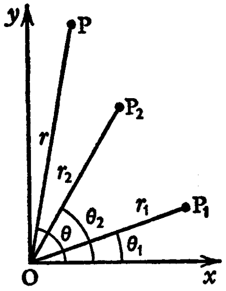


FIG. 2.5c

the vectors OP_1 and OP_2 multiply to the vector OP (for $z=z_1\times z_2$) where OP has absolute value $r=r_1\times r_2$ and argument $\theta=\theta_1+\theta_2$. Multiplication combines expansion and rotation of vectors: the lengths multiply and the angles add (Fig. 2.5*c*).

To return to algebra, we need a more convenient notation than $z=(x, y)$. Write (x, y) as xpy, where p is a 'place-marker' designed to show that x comes first and y second. The object now is to get a suitable notation for p. In the light of what we have said, we expect p to be '$+i$'. But we must get this from the rules (2) which appear in the present notation:

$$\left.\begin{aligned} x_1py_1+x_2py_2&=(x_1+x_2)p\,(y_1+y_2)\\ x_1py_1\times x_2py_2&=(x_1x_2-y_1y_2)p\,(x_1y_2+x_2y_1)\end{aligned}\right\}\quad\ldots\ldots\ldots\ldots(3)$$

Pick out the complex number $1p0$ (vector OA) and write it: $1p0=1$. This serves as *unity* since (3) gives: $1p0\times xpy=xpy$.

Next pick out the complex number $(-1)p0$ (vector OA') and write it: $(-1)p0=-1$. This is in order since, by (3):

$$(-1)^2=(-1)\times(-1)=(-1)p0\times(-1)p0=1p0=1.$$

Further, pick out the complex number $0p1$ (vector OB) and write it: $0p1=i$. Then by (3) i is such that:

$$i^2=i\times i=0p1\times 0p1=(-1)p0=-1.$$

Finally, if x is any real number, write $xp0=x$. From (3):

$$\begin{aligned} x_1p0+x_2p0&=(x_1+x_2)p0=x_1+x_2\\ x_1p0\times x_2p0&=(x_1x_2)p0=x_1x_2.\end{aligned}$$

Hence, $xp0$ is associated with the variable real number x; these numbers add and multiply amongst themselves and they are given by points on the 'real' axis Ox.

The appropriate way of writing p is now apparent. By applications of (3):

$$xpy=xp0+0py=xp0+0p1\times yp0=x+i\times y$$

in terms of the notations adopted. Hence $xpy=x+iy$ and p is '$+i$'.

NOTATION: *The complex number* $z=(x, y)=x+iy$ *where* $+$ *is addition and* i *is such that* $i^2=-1$.

It follows immediately that the rules (2) or (3) can be forgotten, replaced by the ordinary operations of + and × for real numbers, subject only to $i^2 = -1$:

$$\begin{aligned}(x_1+iy_1)+(x_2+iy_2)&=(x_1+x_2)+i(y_1+y_2)\\(x_1+iy_1)\times(x_2+iy_2)&=x_1x_2+i(x_1y_2+x_2y_1)+i^2y_1y_2\\&=(x_1x_2-y_1y_2)+i(x_1y_2+x_2y_1).\end{aligned}$$

We are back where we were with (1). All the operational rules of 2.2 are obeyed by $x+iy$. The set C of complex numbers $z=x+iy$ is a field.

It is established terminology that x is the *real part* and iy the *imaginary part* of the complex number $x+iy$. Such a use of the word 'imaginary', like the use of 'complex' in complex number, scarcely does the concept of a complex number full justice. However, the terms are quite convenient. In particular, the two conditions for the equality of complex numbers appear:

Equation of real and imaginary parts: if $x_1+iy_1=x_2+iy_2$ then:

$$x_1=x_2 \quad\text{and}\quad y_1=y_2.$$

In conclusion, notice a property of i:

$$i(x+iy)=ix+i^2y=-y+ix.$$

Hence, in Fig. 2.5*d*, if P is $z=x+iy$, and Q is $iz=-y+ix$, then OP and OQ are at right angles. In an Argand Diagram, multiplication by i corresponds to rotation through a right angle. For example: $A\,(z=1)$, $B\,(z=i)$, $A'\,(z=i^2=-1)$ and $B'\,(z=i^3=-i)$ are so related.

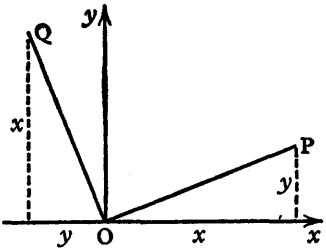


FIG. 2.5*d*

There are now three important interpretations of i:

(1) i is a number such that $i^2=-1$, a root of $x^2+1=0$

(2) i is the vector OB in the Argand Diagram, the unit vector at right angles to the 'real' axis Ox.

(3) 'multiplication by i' is equivalent to 'rotation through a right angle' in the Argand Diagram.

2.6. Integers. We have jobbed forwards from rationals to real and complex numbers, keeping throughout to a system of numbers which is a field, satisfying all the required operational rules. Looking back, we may well be surprised at how much we have left undefined at the outset. In what sense are rationals subject to sums, products and ordering? It is time to job backwards and find out.

The basic entities behind the rationals are the natural numbers, or positive integers, which we get essentially from counting on our fingers. Our starting point is the set J^+ of all positive integers. We still have a choice. *Either* we take 'positive integer' as a primitive (undefined) concept, subject to a set of axioms, and proceed to define sums, products and order in such a way that we derive the properties we desire for the set J^+. How this may be done is shown in 15.1. *Or* we take the positive integers for granted, and simply specify the properties we assume they have. This is the line now followed.

The set J^+ of *positive integers* is $\{1, 2, 3, 4, \ldots\}$. Unspecified positive integers are denoted: $m, n, p, q, r, \ldots$. Their properties, laid out carefully and specifically, are a formidable batch indeed:

(1) There are in J^+ symmetrical operations of *addition* and *multiplication*, specified by addition and multiplication tables of which the essential parts are:

+	1	2	3	4	5	6	7	8	9
1	2	3	4	5	6	7	8	9	10
2	3	4	5	6	7	8	9	10	11
3	4	5	6	7	8	9	10	11	12
4	5	6	7	8	9	10	11	12	13
5	6	7	8	9	10	11	12	13	14
6	7	8	9	10	11	12	13	14	15
7	8	9	10	11	12	13	14	15	16
8	9	10	11	12	13	14	15	16	17
9	10	11	12	13	14	15	16	17	18

×	1	2	3	4	5	6	7	8	9
1	1	2	3	4	5	6	7	8	9
2	2	4	6	8	10	12	14	16	18
3	3	6	9	12	15	18	21	24	27
4	4	8	12	16	20	24	28	32	36
5	5	10	15	20	25	30	35	40	45
6	6	12	18	24	30	36	42	48	54
7	7	14	21	28	35	42	49	56	63
8	8	16	24	32	40	48	56	64	72
9	9	18	27	36	45	54	63	72	81

These are the tables learnt, so painfully, at school. The symmetry is evident, since in each 'matrix' table the entries below the leading downward diagonal are the same as those above.

(2) J^+ is *closed* with respect to the *commutative* operations of addition and multiplication, i.e. if m and n belong to J^+, then $m+n=n+m$ and $mn=nm$ belong to J^+.

(3) Addition and multiplication are both *associative*, i.e. if p, q and r belong to J^+, then $p+(q+r)=(p+q)+r$ and $p(qr)=(pq)r$.

(4) Addition and multiplication are *distributive*, i.e. if p, q and r belong to J^+, then $p(q+r)=pq+pr$ and $(p+q)r=pr+qr$.

(5) J^+ contains an *identity* for multiplication, i.e. unity 1 so that $m \times 1 = 1 \times m = m$ for any m of J^+.

(6) *Cancellation* is valid, i.e. if $m + p = m + q$ then $p = q$
and if $mp = mq$ then $p = q$.

(7) If m and n are different members of J^+ ($m \neq n$), then the *difference* of m and n can be defined as belonging to J^+:

Either $m + p = n$ for some p in J^+ so that the difference $p = n - m$.

Or $m = n + p$ for some p in J^+ so that the difference $p = m - n$.

(8) If a set contains 1, and contains $(n + 1)$ whenever it contains n, then it comprises all members of J^+: the *principle of mathematical induction.*

Apart from the first, defining sums and products, the properties effectively amount to the following. Properties (2)–(6) inclusive specify how many of the operational rules of 2.2 hold for positive integers. They all hold *except*:

Addition: no zero and no negatives.

Multiplication: though there is a unity, there are no reciprocals. Note that the *commutative* rule 3 is a consequence of the *symmetrical* definition of sums and products. Also, though rule 5 (inverses) does not hold, the weaker form of rule 5*A* does hold and *cancellation* is valid. Further, in the absence of inverses, the processes of subtraction and division do not apply; in general, it is useless to try to subtract or divide two positive integers. This is perfectly well-known: 5 pennies cannot be taken from someone with only 3. However, property (7) provides some substitute for subtraction, since m and n have a difference for any m and n ($m \neq n$).

This leads to the last point, covered by the very powerful property (8): what is the order of the integers? Can we write them in sequence: 1, 2, 3, ... n, $(n+1)$, ...? We can, by induction from property (8). Starting with 1, suppose we have proceeded as far as n; then by induction the next integer is $(n + 1)$. The order of J^+ is: add unity 1 to any integer to get the next in the series. This order implies that we always know which of two different integers is the earlier and which is the later in the sequence. Hence we *define*: m is *less* than n ($m < n$) and n is *greater* than m ($n > m$) as two equivalent ways of stating that m is earlier in the sequence of J^+ than n. The difference property (7) now appears: $m<n$ means that a positive integer p exists so that $m+p=n$, i.e. the difference $(n-m)$ is a positive integer.

Hence, J^+ is an ordered set, arranged in sequence 1, 2, 3, ... n, $(n+1)$, ... On the other hand, J^+ is a good deal short of satisfying all the operational rules of algebra; it is *not* a field.

The principle of *mathematical induction* can be elaborated into a powerful method of proof. Mathematical induction is applied in the following form. A property $P(n)$ is to be established, i.e. proved for all positive integers n. Then (i) check that it holds for $n=1$, i.e. $P(1)$ true; (ii) prove that, if it holds for n, then it holds for $(n+1)$, i.e. if $P(n)$ then $P(n+1)$; finally (iii) say that, by induction, $P(n)$ is true for all n. The last step simply means: since $P(1)$ true, so is $P(2)$; since $P(2)$ true, so is $P(3)$; and so on. This is like climbing a ladder: (i) get your foot on the bottom rung; (ii) see how to move a foot from one rung to the next; then (iii) you can climb up the ladder as far as you want.

This simple concept not only provides a ready method of proof; it is also involved in our intuitive way of getting results. We *guess* what a property $P(n)$ should be first by working out the simplest cases, $n=1$, 2, 3, ..., and then by having a shot at a generalisation. We *prove* $P(n)$ by induction. For example: Find the sum of the first n odd integers, i.e. $P(n)=1+3+5+\ldots+(2n-1)$. (i) Try $n=1$: $P(1)=1$; $n=2$: $P(1)=1+3=4$; $n=3$: $P(3)=1+3+5=9$. This suggests $P(n)=n^2$, verified for $n=1$, 2, 3. (ii) If $P(n)=n^2$, then $P\ (n+1)=1+3+5+\ldots+(2n-1)+(2n+1)=n^2+(2n+1)=(n+1)^2$. (iii) So, by induction, $P(n)=n^2$ for all positive integers n.

Two steps now need to be taken, starting from J^+, the positive integers. First, J^+ must be extended to the wider set J of all integers, positive, zero and negative. Then J must itself be extended to the still wider set R of all rationals. We need spend little time on the first step,* which can be regarded as carrying the order of J^+ backwards from 1 to 0, to -1, to -2, ..., adding a new integer each time. There is no positive integer 0 such that $n+0=0+n=n$, so define a new number zero for the purpose. So, $0+1=1$, i.e. 1 is the successor in the order to the new number 0. Next, define a new number (-1) so that

* The formal development of this step is more complicated and it can best be expressed (like the other step) by a construction of pairs of positive integers. This is indicated in 15.1.

$(-1)+1=0$, i.e. 0 is the successor to the new number -1 and (-1) is the negative of 1. Then, define a further number (-2) so that $(-2)+1=-1$, i.e. -1 is the successor to the new number -2. It also follows that $(-2)+1+1=-1+1$, adding 1 to each side, i.e. $-2+2=0$ and (-2) is the negative of 2. The process continues, in the order of the positive integers, producing the opposite series of negative integers. The set J of all integers, still ordered by greater/less, is the result. It differs from J^+ by having a zero and negatives, i.e. rules 4 and 5 for addition now hold. The only lack in J is that there are generally no reciprocals, and hence no division process. A set with this one defect (rule 5 for multiplication not valid, but rule 5*A* on cancellation valid) is called an *integral domain*. J is an integral domain, coming close to the requirements for a field, but failing to make it because of the lack of reciprocals.

The second step, from the set of integers J to that of rationals R, corrects this; it fills in the gap by defining reciprocals and ratios of integers. The construction involved is that of number pairs and 'place-markers' already used with success for complex numbers:

DEFINITION: *The set R of* **rational numbers** *comprises all ordered pairs of integers (p, q), where p and q range over J, $q \neq 0$. Sums and products are:*

$$(p_1, q_1)+(p_2, q_2)=(p_1q_2+p_2q_1, q_1q_2)$$
$$(p_1, q_1)\times(p_2, q_2)=(p_1p_2, q_1q_2).$$

As with complex numbers, the reason for choosing these particular addition and multiplication processes is seen by looking ahead to the final outcome. We have in mind to replace the integer pair (p, q) by pPq where P is a 'place-marker', which in its turn we wish to denote by $\div$ or /, to get the rational number $p\div q$ or p/q. The sum and product rules given achieve this, since we aim at:

$$\frac{p_1}{q_1}+\frac{p_2}{q_2}=\frac{p_1q_2+p_2q_1}{q_1q_2} \quad \text{and} \quad \frac{p_1}{q_1}\times\frac{p_2}{q_2}=\frac{p_1p_2}{q_1q_2}$$

as in elementary algebra. It is interesting to note that the addition process is here the more complicated; for complex numbers it is the product which is involved.

As in the development of the notation $x+iy$ for a complex number, the definition and rules for sums and products of rational numbers justify writing the place-marker P as $\div$ or /. A rational number is then p/q, where p and q are integers ($q \neq 0$), and all the operational rules of 2.2 are obeyed (see 2.9 Ex. 7).

The result is that we justify the original assumption that the set R of rationals is a field. It has been 'manufactured' from the simpler set J of integers (which is not a field since it lacks reciprocals and division) by a process which first pairs off integers (p, q) and then associates them with quotients p/q. R can be called the *quotient field* of J. It is a formalised version of what is done in practice in arithmetic. It is also a general process which can be applied in the construction of other fields, e.g. for polynomials in Chapter 3.

The ordered property of the rationals R, as a reflection of the basic ordering of the integers J, has still to be established. This is best done initially for the set R^+ of *positive rationals* obtained from the set J^+ of *positive integers*, i.e. R^+ is that part of R which corresponds to J^+ as part of J, and a positive rational is a pair (p, q) of positive integers from J^+. Duplication in R^+ is first eliminated by amalgamating 'equivalent' rationals:

$$(p_1, q_1) = (p_2, q_2) \quad \text{if} \quad p_1q_2 = p_2q_1$$

which simply means $p_1/q_1 = p_2/q_2$. Having got rid of this complication, e.g. by confining (p, q) or p/q to relatively prime p and q, we *define*:

$$(p_1, q_1) < (p_2, q_2) \quad \text{if} \quad p_1q_2 < p_2q_1$$

a condition which depends only on the ordering of positive integers (here p_1q_2 and p_2q_1). In terms of the quotient notation:

$$\frac{p_1}{q_1} < \frac{p_2}{q_2} \quad \text{if} \quad p_1q_2 < p_2q_1 \quad \text{by cross multiplication.}$$

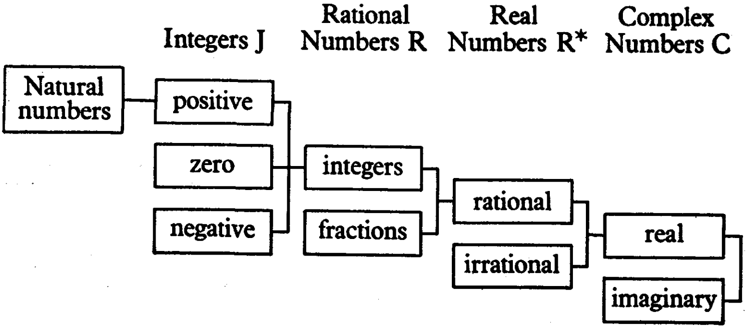
The ordering of zero and negative rationals can then be dealt with. First, zero is less than all positive rationals; then, negative rationals are ordered by:

$$-\frac{p_2}{q_2} < -\frac{p_1}{q_1} < 0 \quad \text{if} \quad p_1q_2 < p_2q_1$$

for any positive integers p_1, p_2, q_1 and q_2.

One conceptual (but genuine) difficulty has been glossed over. Each extension, from the set J^+, to the integral domain J, and to the fields R, R^* and C, involves a widening of the scope of the number system, in such a way that each set is contained within the later ones. The positive integers of J^+ are reproduced with others (zero and negative integers) in J; the integers of J are reproduced with others (fractions p/q, $q \neq 0$ or 1) in R; and so on. The difficulty is that one number appears in various disguises. The positive integer 3 in J^+ is also $+3$ in J, $+\frac{3}{1}$ in R, the real number 3 in R^* and the complex number 3 in C. All we say here is that, though these are all different, they are essentially equivalent, so that we can switch from one to the other as necessary. There is, however, rather more to be said on the matter (see 7.4 below).

The build-up of the number system can be shown in summary in the classificatory scheme:



2.7. Finite sets of integers. The set of positive integers is the number concept derived from counting on the fingers; but the derivation is not immediate. The set of positive integers, and the more developed sets of numbers, are all infinite sets. To one who uses only his fingers for counting, the infinite sequence of positive integers must be quite a sophisticated idea. He cannot have much idea of integers in thousands or millions, or even of the eleven or twelve times tables. He would start with 0 on the thumb of his left hand, and proceed 1, 2, 3, ... until he reaches 9 on the thumb of his right hand (or he might go from 1 to 10, much the same thing). But then he goes back to 0 again on the thumb of his left hand and starts a new cycle to 9. Suppose, however, he has been introduced to a clock and is keeping tabs on the hours. Looking at the clock face, he would tick off the hours from 0 to 1, 2, 3, ... until he comes to 11. He would then go back to 0 and start again. So, for example, in counting 78 sticks, he would go 7 complete rounds and find himself left with 8 on his hands. Equally, 78 hours after midnight on D-day, he would have 6 on the clock, the hour-hand having done 6 complete revolutions. If he counts in tens, the number which is really 78 would appear to him as 8. If he counts in twelves, as in time-keeping, the same number would appear as 6.

There is an important idea here. In a count on two hands, only the ten integers $\{0, 1, 2, \dots 9\}$ occur and each number is replaced by its residue or remainder on division by 10. Similarly, in time-keeping with a clock, every hour shows up as the remainder after division by 12, only the twelve integers $\{0, 1, 2, \dots 11\}$ being used. This suggests that we should investigate what happens if, instead of the whole set of integers (positive, zero and negative), we keep only the remainder on division by a selected positive integer n:

Definition: *The* **integers modulo n** $\{0, 1, 2, \dots (n-1)\}$ *(mod n) are the finite set obtained by keeping only the remainder on dividing an integer by n.*

For example, ordinary counting is mod 10, time-keeping the hours is mod 12.

Remember the properties of the *integral domain* J: all the operational rules of algebra hold *except* that there are no reciprocals, but cancellation is valid:

$$\text{If} \quad mp = mq \ (m \neq 0), \quad \text{then } p = q.$$

As one particular case ($q = 0$). the rule is:

$$\text{If} \quad mp = 0, \quad \text{then either } m = 0, \text{ or } p = 0, \text{ or both.}$$

There are no *divisors of zero*; no two (non-zero) integers multiply to zero.

Our object is to see whether the finite set of integers (mod n) does better or worse than J in respect of the operational rules obeyed. Start with the integers modulo 5: $\{0, 1, 2, 3, 4\}$ (mod 5), corresponding to counting on one hand of five fingers, only the remainder on division by 5 being retained. The operations of addition and multiplication are, otherwise, exactly as in ordinary arithmetic. Consider 3 and 4 (mod 5):

$$3 + 4 = 7 = 1 \times 5 + 2 \qquad \text{replaced by } 2$$

$$3 \times 4 = 12 = 2 \times 5 + 2 \qquad \text{again replaced by } 2$$

$$\text{i.e.} \qquad 3 + 4 = 2 \quad \text{and} \quad 3 \times 4 = 2 \quad (\text{mod } 5).$$

Other examples are: $2 + 3 = 0$; $1 + 4 = 0$; $2 \times 3 = 1$; $1 \times 4 = 4$.

Hence we build up the addition and multiplication tables for integers (mod 5):

$+$	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

$\times$	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

For the arithmetic of integers (mod 5), nothing more is required; these tables *define* the operations of $+$ and $\times$. It remains to check all the operational rules of 2.2. In doing so, we cannot fail to notice one very helpful property of the tables: each integer appears exactly once in every row or column (ignoring 0 for multiplication).

The definitions of $+$ and $\times$ are symmetrical and the result is always an integer of the set $\{0, 1, 2, 3, 4\}$ (mod 5), i.e. the tables are symmetrical about the leading (downward) diagonal and contain only the five integers. Hence the set is *closed* and *commutative*. It is also *associative*, e.g. $2+(3+4)=2+2=4$ and $(2+3)+4=0+4=4$ and similarly for products. The *distributive* law holds, e.g.

$$2(3+4)=2\times 2=4 \quad \text{and} \quad 2\times 3+2\times 4=1+3=4.$$

We must look very carefully for identities and inverses. For addition, the *zero* is 0, since an integer is unaltered by addition of 0. If p is an integer, its negative $(-p)$ is such that $p+(-p)=0$. Hence, we look for 0 in any row of the addition table, and every row has a zero. So *negatives* exist:

$$(-1)=4; \quad (-2)=3; \quad (-3)=2 \quad \text{and} \quad (-4)=1.$$

These can be checked by the remainders on division by 5:

$$(-1)=(-1)\times 5+4; \quad (-2)=(-1)\times 5+3; \quad (-3)=(-1)\times 5+2;$$
$$(-4)=(-1)\times 5+1.$$

With negatives defined, subtraction follows, e.g.

$$2-4=2+(-4)=2+1=3.$$

For multiplication, the *unity* is 1, since an integer is unaltered on multiplication by 1. The reciprocal p^{-1} of any integer p is given by $p\times p^{-1}=1\ (p\neq 0)$. We look for 1 in a row of the multiplication table, and every row has a 1. So:

$$2^{-1}=3; \quad 3^{-1}=2; \quad 4^{-1}=4$$

together with the fact (always true) that 1 is its own reciprocal. Hence *reciprocals* exist and so does division to give p/q $(q \neq 0)$. For example:

$$\tfrac{3}{2} = 3 \times \tfrac{1}{2} = 3 \times 2^{-1} = 3 \times 3 = 4; \quad \tfrac{4}{2} = 4 \times \tfrac{1}{2} = 4 \times 2^{-1} = 4 \times 3 = 2.$$

Hence the integers (mod 5) have everything. Manipulation of the five integers appears strange at first, but it satisfies all the operational rules. The integers (mod 5) are a *field*. They are an improvement on the set of all integers; they have reciprocals and division (see 2.9 Ex. 19).

This may be an accidental result, connected with the fact that we have selected 5, a prime, to start with. It is necessary to try again. Consider the integers, modulo 4 and modulo 6. These would arise if we counted on one hand, with 4 or 6 fingers. The addition tables look very much like that above for mod 5, just one row and column more or less. The integers mod 4 or mod 6 are well-behaved for sums (2.9 Ex. 21). It is the multiplication table that varies:

{0, 1, 2, 3} (mod 4)

×	0	1	2	3
0	0	0	0	0
1	0	1	2	3
2	0	2	0	2
3	0	3	2	1

{0, 1, 2, 3, 4, 5} (mod 6)

×	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

The pattern of these tables is different from that for integers (mod 5). In particular, some rows do not contain the integer 1; instead they have extra zeros. The effect is to be seen in the definition of reciprocals. Unity is still 1 and 1 is the reciprocal of itself. For *integers* (*mod* 4), $3^{-1} = 3$ so that 3 is also the reciprocal of itself; but there is no integer which multiplies 2 to give 1, i.e. 2 has no reciprocal. Rule 5 on reciprocals breaks down. Worse still, the weaker rule 5*A* also fails. It is seen that $2 \times 2 = 0$, i.e. 2 is a divisor of zero. This arises because $4 = 2 \times 2$ is not a prime; when 4 is replaced by 0 (mod 4), then $2 \times 2 = 0$. Equally, for *integers* (*mod* 6): $6 = 2 \times 3$, and there are difficulties with 2, 3 and 4. From the table, $5^{-1} = 5$ so that 5 (like 1) is the reciprocal of itself. But there are no reciprocals for 2, 3 or 4. Moreover these are divisors of zero: $2 \times 3 = 3 \times 4 = 0$.

The conclusion is that integers (mod 5) form a field; they are better behaved than J itself. On the other hand, integers (mod 4) and integers (mod 6) satisfy neither rule 5 nor rule 5A; they have zero divisors and are worse than J. The result is, in fact, quite general: integers (mod n) are a field if n is prime, but fail to satisfy rules 5 and 5A if n is not prime.

2.8. The binary system. The simplest finite set is that comprising only the integers 0 and 1, the zero for addition being 0 and the unity for multiplication being 1. All that is needed for the set to be a field is that 1 should be both the negative and the reciprocal of itself. This is achieved with the set {0, 1} (mod 2).

It is useful, however, to start more simply. In the set {0, 1}, assume that sums and products have the following three (very familiar) properties

(i) Addition of 0 to any integer leaves it unchanged.

(ii) Multiplication of any integer by 0 gives 0.

(iii) Multiplication of any integer by 1 leaves it unchanged.

The addition and multiplication tables for {0, 1} are then:

+	0	1
0	0	1
1	1	*

×	0	1
0	0	0
1	0	1

and the only thing left open is the meaning of $1+1$ to fill the space marked *. There are several possibilities, still using only 0 and 1:

(1) Specify $1+1=0$ so that:

+	0	1
0	0	1
1	1	0

The set is now {0, 1} (mod 2). Sums and products are the usual arithmetical ones, provided only that every integer is replaced by its remainder on division by 2, e.g. $1+1=2$ replaced by 0. The set is a field, the simplest instance of the field {0, 1} (mod n), n prime.

One application is in the treatment of even and odd integers, and so of other entities which can be described as even and odd. An even integer has 0, and an odd integer has 1, as remainder on division by 2. So all even integers appear as 0 and all odd integers as 1 (mod 2). The tables for sums and products then state:

Addition: even + even = odd + odd = even and even + odd = odd.
Multiplication: even × even = even × odd = even and
odd × odd = odd.

(2) Specify $1+1=1$ so that:

+	0	1
0	0	1
1	1	1

The set is now a different one and, since 1 has no negative, it is not a field. One interpretation which can be given to addition here is:

$$p+q = \text{larger of } (p, q).$$

The difference as compared with ordinary arithmetic is that $1+1=1$ and so

$$1+1+\ldots(n \text{ times}) = 1 \quad (\text{and } \textit{not } n).$$

An application of such a set appears in 4.4 below.

(3) Allow now for numbers with two (or more) digits, each digit being either 0 or 1. For two digit numbers, 00 and 01 are merely 0 and 1 in disguise, but 10 and 11 are new. Specify $1+1=10$ so that:

+	0	1
0	0	1
1	1	10

Suppose, further, that the simple arithmetic of multi-digit numbers still applies.

So, for *addition:*

$$\begin{array}{r} 10 \\ 1 \\ \hline 11 \end{array} \qquad \begin{array}{r} 10 \\ 10 \\ \hline 100 \end{array} \qquad \begin{array}{r} 10 \\ 11 \\ \hline 101 \end{array} \qquad \begin{array}{r} 11 \\ 11 \\ \hline 110 \end{array}$$

where $1+1=10$ and a 'carry one' process is required, e.g. in adding

11 and 11, the right-hand 1's add to 10, the 1 is carried forward to the next digit:

$$1+1+1=1+(1+1)=1+10=11.$$

For *multiplication*:

$$\begin{array}{r}10\\1\\\hline 10\end{array}\qquad \begin{array}{r}10\\10\\\hline 100\end{array}\qquad \begin{array}{r}10\\11\\\hline 10\\10\\\hline 110\end{array}\qquad \begin{array}{r}11\\11\\\hline 11\\11\\\hline 1001\end{array}$$

with the same 'carry one' process. This is the *binary system* for handling integers.

Another approach makes the nature of the binary system clear. In the *decimal system*, any integer is written in terms of powers of 10. Reading the number from right to left, the first digit represents so many (0, 1, 2, ... or 9) units, the second so many 10's, the third so many 100's ($100=10^2$), and so on. For example:

$$147=1\times 10^2+4\times 10+7 \quad \text{(equals 7 in mod 10).}$$

Generally:
$$a_n a_{n-1}\ldots a_1 a_0=\sum_{r=0}^{n} a_r 10^r$$
$$=a_n 10^n+a_{n-1}10^{n-1}+\ldots+a_1 10+a_0$$

where the a's are from the set $\{0, 1, 2, \ldots 9\}$.

In the *binary system*, 2 replaces 10. Any integer is written in terms of powers of 2 and the digits (multiples) are written 0 or 1. From right to left, the first digit is 0 or 1, the second is a multiple (0 or 1) of 2, the third a similar multiple of $2^2=4$, and so on. Successive powers of 2 are:

n	1	2	3	4	5	6	7	8	...
2^n	2	4	8	16	32	64	128	256	...

Hence 147 must start with $1\times 2^7=128$ and remainder 19:

$$147=1\times 2^7+19; \quad 19=1\times 2^4+3; \quad 3=1\times 2+1.$$

So: $147=1\times 2^7+0\times 2^6+0\times 2^5+1\times 2^4+0\times 2^3+0\times 2^2+1\times 2+1$

i.e. $147=10010011$ taking 8 digits altogether.

The integers up to 32 appear in the binary notation:

Decimal	*Binary*	*Decimal*	*Binary*	*Decimal*	*Binary*	*Decimal*	*Binary*
1	1	9	1001	17	10001	25	11001
2	10	10	1010	18	10010	26	11010
3	11	11	1011	19	10011	27	11011
4	100	12	1100	20	10100	28	11100
5	101	13	1101	21	10101	29	11101
6	110	14	1110	22	10110	30	11110
7	111	15	1111	23	10111	31	11111
8	1000	16	10000	24	11000	32	100000

In general: $$b_n b_{n-1} \dots b_1 b_0 = \sum_{r=0}^{n} b_r 2^r$$
$$= b_n 2^n + b_{n-1} 2^{n-1} + \dots + b_1 2 + b_0$$

where the b's are from the set $\{0, 1\}$.

As a check, multiplication of 11×11 gives 1001 in the binary system (as above). Here 11 is 3 and 1001 is 9; the product is $3 \times 3 = 9$.

The binary system of denoting numbers has entered the popular domain since the introduction of high-speed computers. Electronic circuits are well-adapted to handling only two digits 0 or 1 (circuits open or closed). There are other possible systems, e.g. the duodecimal based on 12 and the octal based on 8. Numbers in octals are easily linked to the binary system used in computers (see 2.9 Ex. 25).

2.9. Exercises

1. Illustrate a difficulty with recurring decimals by adding $4/3 = 1{\cdot}\dot{3}$ to $5/3 = 1{\cdot}\dot{6}$ to give $3 = 2{\cdot}\dot{9}$. Is there any difference between $0{\cdot}\dot{9}$ and 1?

2. *Rationals as decimals.* Consider the rational $\frac{p}{q}$ where p and q are integers, $q > 1$, p and q relatively prime. Show that $\frac{p}{q}$ is a terminating decimal if and only if q has only 2's and 5's as factors, and that the decimal recurs otherwise. Illustrate by writing 1/3, 1/7, 1/11 and 1/13 as recurring decimals. An approximation to π is $22/7 = 3{\cdot}\dot{1}4285\dot{7}$ and 355/113 is closer. Use 355/113 to illustrate the problem of finding whether a decimal recurs.

3. Attempt to make subtraction $a - b$ a basic operation in the set of rationals and check that the operational rules of 2.2 do not all hold. In particular, show

that the associative rule fails: $a-(b-c) \neq (a-b)-c$. Illustrate the subsidiary nature of division $a \div b$ similarly.

4. Show that the operation 'to the power' ($a^b = a$ to the power b) is not associative in the set of rationals, i.e. a to the power (b to the power c) is not the same as (a to the power b) to the power c. See 1.9 Ex. 17.

5. *Pythagoras' Theorem.* Triangle ABC is right-angled at C; a, b and c are the lengths of the sides. Then $c^2 = a^2 + b^2$. Given rational a and b, to find c is equivalent to solving $x^2 - k = 0$, where the rational $k = a^2 + b^2 > 0$. Show that c is irrational except in such special cases as $a = 3$, $b = 4$.

6. *$R(\sqrt{2})$ and $R(i)$ as fields.* The set of $\alpha = a + b\sqrt{2}$ (a and b rationals) satisfies all the operational rules of 2.2 with $+$ and $\times$ defined by (1) of 2.3. Check the rules for $+$, noting that zero is 0 ($\alpha = 0 + 0 \times \sqrt{2}$) and that $-\alpha = (-a) + (-b)\sqrt{2}$. For the rules for $\times$, see 2.3. Check the distributive rule: $\alpha(\beta + \gamma) = \alpha\beta + \alpha\gamma$ by substituting full expressions for α, β and γ and amplifying both sides. Similarly show that the set of $z = a + ib$ (a and b real) is a field under the $+$ and $\times$ rules of (1) of 2.5.

7. *The field of rationals.* Show that the set of rationals $\alpha = (p, q)$, for p and q integers, satisfies all the operational rules of 2.2 with $+$ and $\times$ defined as in 2.6. Note that zero is $(0, q)$ for any $q \neq 0$ and that $-\alpha = (-p, q)$; that unity is (p, p) for any $p \neq 0$ and that $\alpha^{-1} = (q, p)$.

8. *The irrational π.* The area of a circle of unit radius is π, approximated (as in Euclid) by the area of inscribed and circumscribed (regular) polygons. For pentagons (Fig. 2.9), with $OA = OR = OB = 1$, write:

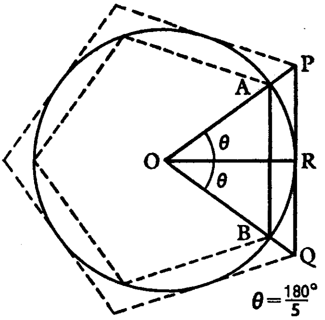


FIG. 2.9

$$\text{Area } OAB = \tfrac{1}{2}OA\,.\,OB \sin \angle AOB = \tfrac{1}{2}\sin\frac{360°}{5}$$

$$\text{Area } OPQ = OR\,.\,PR = OR^2 \tan \angle POR = \tan\frac{180°}{5}.$$

Hence show: $5\tan\dfrac{180°}{5} < \pi < \dfrac{5}{2}\sin\dfrac{360°}{5}$ and evaluate these bounds from trigonometric tables. Generalise and obtain a sequence of nested intervals defining π: $\left[n\tan\dfrac{180°}{n}, \dfrac{n}{2}\sin\dfrac{360°}{n}\right]$ for $n = 3, 4, 5, 6, \ldots$.

9. Consider the set S of *rationals* x such that $x^2 > 2$. In the field of rationals, S has a lower bound (e.g. $x = 1$). Suppose a (rational) is any lower bound, so that $a^2 < 2$. Show that there is a larger rational $b > a$ so that $b^2 < 2$ (e.g. by expressing a^2 as a decimal, short of 2). Now look at the same set S of rationals, but in the field of real numbers. Then $a = \sqrt{2}$ is such that no real number can be inserted between a and the rationals of S. Hence deduce that S has no *rational* GLB but that it has the *real* GLB $\sqrt{2}$.

10. Attempt to define i as dividing L and G in a cut of R, where L comprises rationals x such that $x^2+1<0$ and G rationals x such that $x^2+1>0$. Show that this fails because L is empty, i.e. i is not real.

11. In Fig. 2.5*b*, show that P is (x_1+x_2, y_1+y_2) where P_1 is (x_1, y_1) and P_2 is (x_2, y_2). Deduce that $z=z_1+z_2$ is given by OP.

12. Show that the general complex number can be written

$$z=r(\cos\theta+i\sin\theta)$$

where r is the absolute value and θ the argument (see Fig. 2.5*a*). Show that $r=1$ for each of the four points A, B, A' and B' of an Argand Diagram and that $\theta=0°$, $90°$, $180°$ and $270°$, giving $z=1$, i, -1 and $-i$ respectively.

13. Show that $z_1=r_1(\cos\theta_1+i\sin\theta_1)$ and $z_2=r_2(\cos\theta_2+i\sin\theta_2)$ give:

$$z_1z_2=r_1r_2\{\cos(\theta_1+\theta_2)+i\sin(\theta_1+\theta_2)\}$$

using the addition formulae (Appendix A.7). Interpret in Fig. 2.5*c*.

14. *De Moivre's Theorem.* If $z=r(\cos\theta+i\sin\theta)$, then

$$z^2=r^2(\cos 2\theta+i\sin 2\theta).$$

Generalise and deduce the theorem of de Moivre (1667–1764):

$$(\cos\theta+i\sin\theta)^n=\cos n\theta+i\sin n\theta \quad (n \text{ a positive integer}).$$

15. Show that the square of $(1+i)$ is $2i$ and deduce that $\sqrt{i}=\pm(1+i)/\sqrt{2}$.

*16. *Powers of a complex variable.* Consider z^a where $z=r(\cos\theta+i\sin\theta)$ and a is a rational number. If a is an integer, show that $z^a=r^a(\cos a\theta+i\sin a\theta)$ and write $1/z$ as a particular case. Illustrate the case of fractional a by showing that one value of $\sqrt{z}$ is $\sqrt{r}(\cos\frac{1}{2}\theta+i\sin\frac{1}{2}\theta)$. (Note: square up to z.) Then write $z=r\{\cos(n360°+\theta)+i\sin(n360°+\theta)\}$ for any integer n and show that another value of $\sqrt{z}$ is

$$\sqrt{r}\{\cos(180°+\tfrac{1}{2}\theta)+i\sin(180°+\tfrac{1}{2}\theta)\}=-\sqrt{r}(\cos\tfrac{1}{2}\theta+i\sin\tfrac{1}{2}\theta).$$

Check that these are the only values and that $\sqrt{i}=\pm(1+i)/\sqrt{2}$.

17. *Cube roots of unity.* Write $z_1=\frac{1}{2}(-1+i\sqrt{3})$ and $z_2=\frac{1}{2}(-1-i\sqrt{3})$. Show that $z_1^2=z_2$, $z_1^3=1$; and that $z_2^2=z_1$, $z_2^3=1$. Deduce that z_1 and z_2 are cube roots of 1 (solutions of $x^3-1=0$). What is the third cube root?

18. By mathematical induction, show that: $1+2+3+\ldots+n=\frac{1}{2}n(n+1)$ and that $1+2+2^2+\ldots+2^{n-1}=2^n-1$. Generalise the first by showing that $\frac{1}{2}n\{2a+(n-1)d\}$ is the sum of n terms of an AP (first term a, common difference d), and the second to give $a(r^n-1)/(r-1)$ as the sum of n terms of a GP (first term a, common ratio $r\neq 1$).

19. Contrast the set $\{0, 1, 2, 3, 4\}$ (mod 5) with the set $\{0, 1, 2, 3, \ldots\}$ of all positive integers (including zero). The former as a field is closed under $+$, $-$ $\times$ and $\div$; the latter lacks differences and quotients. Establish that the integers (mod 5) have no primes, e.g. $1=2\times 3=4\times 4$; that powers are defined, e.g. $2^2=4$, $2^3=3$, $2^4=1$; that the only perfect square (apart from 0 and 1) is $4=2\times 2=3\times 3$. Deduce that the roots of $x^2=4$ are $x=2$ or 3 in the field of integers (mod 5) and compare with the solution for all positive integers ($x=2$ only).

20. Construct + and × tables for the integers (mod 3) and show that the set is a field, similar to the integers (mod 5).

21. Show that the addition tables for the integers (mod 4) and (mod 6) are

+	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

+	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

with the same pattern as for the integers (mod 5) of 2.7.

*22. *Quadratics over the field of integers* (*mod* 5). Consider the quadratic ax^2+bx+c where a, b and c are integers; it is known that $ax^2+bx+c=0$ has at most two roots in the set of integers. Illustrate that the same result holds if a, b and c are integers (mod 5) by showing that

$$x^2+4=(x-1)(x-4) \quad \text{and} \quad x^2+2x+2=(x-1)(x-2)$$

are unique factors, and hence that $x^2+4=0$ has roots $x=1$, 4 and that $x^2+2x+2=0$ has roots $x=1, 2$ (mod 5).

*23. Show that the result of Ex. 22 fails when a, b and c are integers (mod 6) by checking that $x^2+x=x(x-5)=(x-2)(x-3)$ so that $x^2+x=0$ has four roots $x=0, 2, 3, 5$ (mod 6).

24. Show that $0 . a_1a_2 \ldots a_n$ is $\sum_{r=1}^{n} a_r 10^{-r}$ in decimals and $\sum_{r=1}^{n} a_r 2^{-r}$ in binary. Show that, in the binary system: $1/8=0{\cdot}001$, $1/4=0{\cdot}01$, $3/8=0{\cdot}011$, $1/2=0{\cdot}1$, $5/8=0{\cdot}101$, $3/4=0{\cdot}11$ and $7/8=0{\cdot}111$.

25. *Octal system.* Count in powers of 8 instead of 2 (binary) or 10 (decimal) and show that any integer appears as:

$$b_n\, b_{n-1} \ldots b_1\, b_0 = \sum_{r=0}^{n} b_r\, 8^r$$

where the b's are from {0, 1, 2, ... 7}. Check that:

Binary	*Octal*	*Decimal*
100010	42	34
111010	72	58
100001010	412	266

are three integers in alternative forms. Notice that 4 is 100 and 2 is 010 in binary and hence that 42 in octal becomes 100010 in binary. Devise a simple translation from octal to binary, each digit in octal becoming a set of three digits in binary.

CHAPTER 3

POLYNOMIALS

3.1. The fundamental theorem of arithmetic. A schoolboy knows that any integer can be factorised into primes. He also knows that the 'highest common factor' or H.C.F. of two integers is to be got by comparing their factors and by picking out the common ones. For example:

$$\left.\begin{aligned}140&=2\times 70=2\times 2\times 35=2^2\times 5\times 7\\1155&=3\times 385=3\times 5\times 77=3\times 5\times 7\times 11\end{aligned}\right\}\ \text{H.C.F.}=5\times 7=35.$$

The usual method adopted is to 'fish around' for factors, i.e. inspection for divisibility by 2, 2^2, ...; then by 3, 3^2, ..., by 5, 5^2, ... and so on through the primes. This is altogether too slow for larger numbers. Surely there is a more systematic approach? More basically important: why is it assumed that any factorisation achieved is unique? It may well be, but surely a proof is needed? A systematic development proceeds as follows.

In the set J^+ of positive integers, 1 has unique properties; it is the identity for multiplication, it has no factors itself and it does not affect the factors of any other integer. Consider the set J^+ $(n>1)$ apart from 1. Let m and n $(m<n)$ be any two integers and divide m into n to give a quotient q_1 and a remainder r_1. Both q_1 and r_1 are integers and $r_1<m$, except that $r_1=0$ is possible (in which case the division process is complete). Suppose $r_1\neq 0$ and proceed by dividing r_1 into m to give quotient q_2 and remainder $r_2<r_1$. Suppose $r_2\neq 0$, divide r_2 into r_1 to give quotient q_3 and remainder $r_3<r_2$. This process continues until a remainder is found equal to zero (division process complete). This must happen sooner or later since the integral remainders are decreasing $(r_1>r_2>r_3>\ldots)$. Let r_k be the last non-zero remainder. It is the H.C.F. of m and n since, jobbing backwards, r_k divides r_{k-1}, divides r_{k-2}, divides r_{k-3}, ... divides r_1, divides m, and divides n. More formally:

$$n=q_1m+r_1;\ m=q_2r_1+r_2;\ r_1=q_3r_2+r_3;\ \ldots$$
$$r_{k-2}=q_kr_{k-1}+r_k;\ r_{k-1}=q_{k+1}r_k.$$

So:
$$\begin{aligned} r_k = r_{k-2} - q_k r_{k-1} &= r_{k-2} - q_k(r_{k-3} - q_{k-1} r_{k-2}) \\ &= -q_k r_{k-3} + (1 + q_k q_{k-1}) r_{k-2} \\ &= \dots \\ &= (\text{integer})m + (\text{integer})n \quad \text{eventually.} \end{aligned}$$

The integers in the last line are positive or negative.

This process of repeated division is the *Division Algorithm*, or Euclid's Algorithm.* It leads to the result:

THEOREM: *If c is the H.C.F. of two positive integers m and n, then there exist positive or negative integers λ and μ so that*

$$\lambda m + \mu n = c \dots\dots\dots\dots (1)$$

Example: $\left.\begin{aligned} 1155 &= 8 \times 140 + 35 \\ 140 &= 4 \times 35 \end{aligned}\right\}$ H.C.F. of 1155 and 140 is 35

and $35 = 1155 - 8 \times 140$ i.e. (1) with $\lambda = 1$ and $\mu = -8$.

Continuing, we say that two positive integers m and n are *relatively prime* if they have no common factors (except 1). Their H.C.F. is 1, and (1) becomes:

If m and n are relatively prime, then λ and μ exist so that

$$\lambda m + \mu n = 1 \dots\dots\dots\dots (2)$$

Further, a positive integer p is *prime* if it has no factors other than 1 and p itself. A prime p is also relatively prime to all integers except 1 and multiples of p. Two results can now be obtained in succession, leading to the fundamental result on factorisation.

THEOREM: *If p is prime and divides mn, where m and n are two positive integers, then either p divides m, or p divides n, or both*.........(3)

Though this is a result used automatically in elementary arithmetic, it still needs proof. If p does divide m, we need go no further. If p does not divide m, then p and m are relatively prime and, by (2), λ and μ exist for:

$$\lambda p + \mu m = 1.$$

So
$$\lambda pn + \mu mn = n.$$

But p divides mn (given) and it divides pn; hence it divides n.

Q.E.D.

* 'Algorithm' is a rule for computation, a modern mis-spelling of 'algorism'. The Latin 'algorismus' derives from the surname of an Arab mathematician.

THEOREM: *Every positive integer n can be factorised into primes:*

$$n = p_1 p_2 \dots p_i \quad \text{for some } i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

If n is prime, it is itself the only prime factor. If n is not prime, then it has factors (other than 1 and n itself): $n = n_1 n_2$ where n_1 and n_2 are both less than n. Now take n_1 and n_2 in turn and proceed with factorisation. The process stops when only primes are left. This must happen sooner or later since the integers are getting smaller at each step. Q.E.D.

It remains to show that the factorisation (4) is unique. Suppose that:

$$n = p_1 p_2 \dots p_i = q_1 q_2 \dots q_j \quad (p\text{'s and } q\text{'s prime}).$$

Then p_1 divides n and so divides $q_1 q_2 \dots q_i$. By (3), p_1 divides at least one of the q's and (by suitable shuffling) it can be taken that p_1 divides q_1. But p_1 and q_1 are primes, so: $p_1 = q_1$. Divide out $p_1 = q_1$ and proceed to treat:

$$p_2 p_3 \dots p_i = q_2 q_3 \dots q_j$$

in the same way. In the end, each p is identified with a q until all are taken. Hence, $i = j$ and the p's and q's are the same set of primes. Hence:

FUNDAMENTAL THEOREM OF ARITHMETIC: *Every positive integer $n(>1)$ can be uniquely factorised into a product of primes:*

$$n = p_1 p_2 \dots p_i.$$

In this result, the integer 1 is excluded since it divides every integer, i.e. it is the *unit* of the set of positive integers. Nothing is changed in $n = p_1 p_2 \dots p_i$ if n or any of the unique factors is multiplied by 1. More generally for a set S:

DEFINITION: *A* **unit** *of S is an element of S which divides (is a factor of) every element of S.*

So, just as 1 is the only unit of the set of positive integers, 2 is the only unit of the set of even positive integers. There may be more than one unit, e.g. the set of all integers (positive and negative) has two units $+1$ and -1.

3.2. Gaussian integers. As a digression of some interest, consider the field C of complex numbers $x + iy$, where x and y are real numbers. Various subsets of C can be taken, e.g. $a + ib$ where a and b are rationals. One interesting subset is that of $m+in$ where m and n are integers from J. The field C is represented by all points in the plane Oxy on an Argand Diagram; the subset consists of the points (m, n), where m and n are integers, forming a lattice (or trellis-work) over the plane as shown in Fig. 3.2. As in 2.3, the subset of $m+in$ is obtained from the integral domain J by adjunction of the single element i. It can be denoted $J(i)$, the set of *Gaussian Integers*, after Gauss (1777–1855). $J(i)$ is an integral domain like J, as is easily established (3.9 Ex. 2). All the operational rules apply, except for reciprocals and division, but including cancellation.

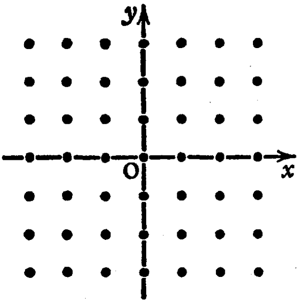


FIG. 3.2

Let us attempt to extend the idea of prime factorisation to $J(i)$. The fundamental theorem shows that this is all right (unique) for positive integers, and for all (positive and negative) integers only a little amendment is needed. Here -1 is a unit as well as 1 and both are ignored in factorising. The process is still unique since any negative integer can be handled:

$$-5=(-1)\ 5=1\ (-1)\ 5$$

and $$-6=(-1)\ 2\ 3=1\ (-1)\ 2\ 3$$

as for positive integers:

$$5=1\ 5=(-1)\ (-1)\ 5$$

and $$6=1\ 2\ 3=(-1)\ (-1)\ 2\ 3.$$

Apart from the units 1 and (-1), factorisation is unique. However, Gaussian integers $m+in$ seem to raise further difficulties. To illustrate, take the prime 5:

$$\begin{aligned}5&=(2+i)(2-i)=(1+2i)(1-2i)\\&=(-2-i)(-2+i)=(-1-2i)(-1+2i).\end{aligned}$$

In $J(i)$, it would appear that 5 is *not* prime, and that factorisation is *not* unique. The first thought is correct; $5=5+i0$ can be split into factors like $(2+i)(2-i)$ in $J(i)$. The second, however, is not correct. Factorisation is still unique in $J(i)$. The point is that there are four *units* which divide every Gaussian integer; they are $1, -1, i, -i$. So:

$(2+i)$; $(-2-i)=(-1)(2+i)$; $(-1+2i)=i(2+i)$; $(1-2i)=-i(2+i)$

are all essentially the same, i.e. $(2+i)$. Hence, apart from the units $(\pm 1, \pm i)$, the factorisation of 5 is unique: $5=(2+i)(2-i)$.

The Gaussian integers have four units $(\pm 1, \pm i)$ but, with this convention, factorisation into primes is unique. The primes in the Gaussian integers $J(i)$ can be different from those in the integers J. Some are the same, e.g. 3 is a prime in both. Some are different, e.g. 5 is prime in J but not in $J(i)$, whereas $(2+i)$ is prime in $J(i)$ but doesn't appear in J. The result has been illustrated here, not proved strictly; but the proof is very similar to that developed for integers in 3.1 above.*

3.3. Polynomials. A boy studying school algebra would not hesitate to say what he means by a 'polynomial'. He might describe it as an expression such as the quadratic $2x^2-x-3$ or the cubic x^3-3x^2+4. Or, if he is a little more pedantic, he might say that the word does double duty, as an adjective and as a noun, so that $2x^2-x-3$ or x^3-3x^2+4 is a polynomial expression, or more shortly a polynomial. If asked to pursue the topic, he might well say that, in application, polynomials are usually equated to zero to give a polynomial equation such as $2x^2-x-3=0$ or $x^3-3x^2+4=0$. The problem is to solve the equation, to find its roots. One way is to factorise the polynomial and deduce the roots. For example: $2x^2-x-3$ has factors $(x+1)(2x-3)$ and the equation $2x^2-x-3=0$ has roots -1 and $\frac{3}{2}$. Again: $x^3-3x^2+4=(x+1)(x-2)^2$ so that $x^3-3x^2+4=0$ has roots -1 and 2 (twice). Conversely, if the roots of the equation are found, the factors of the polynomial follow. For example, the roots of $x^2+x+1=0$ are $-\frac{1}{2}(1\pm i\surd 3)$ by the well-known formula for the quadratic (Appendix A.3), and x^2+x+1 has factors

$$\left(x+\tfrac{1}{2}+i\,\frac{\surd 3}{2}\right)\left(x+\tfrac{1}{2}-i\,\frac{\surd 3}{2}\right).$$

This is, in fact, something of a mess. There are two different uses of 'x' according as we deal with the polynomial expression (function) or with the polynomial equation. In a *polynomial function* such as $y=x^3-3x^2+4$, we have x as a variable with a domain to be specified. For example, if the domain comprises all real numbers, the function fixes a real number y to correspond to each real number x we care to

* See Birkhoff and MacLane: *A Survey of Modern Algebra* (Macmillan, N.Y., Revised Edition, 1953), pp. 413–16.

select. In a *polynomial equation* such as $x^3-3x^2+4=0$, we are interested in the solution set, with one or more values of x which are the roots and which can be regarded as fixed values to be determined. Here there is some lack of precision on what number system is in mind for x. An opportunist attitude is often taken, x being allowed to be rational, real or complex according to what turns up. There is also some uncertainty on what numbers we can use for the coefficients in the polynomial. Does it have integral, rational, real or complex coefficients? We need to say which. This is particularly important in handling the parametric form, e.g. ax^2+bx+c as the general quadratic or ax^3+bx^2+cx+d as the general cubic. The replacement set of the parameters a, b, c, ... must be specified.

We have to dig very deep indeed to get down to a good foundation. It is worth while making an effort here because of the insight gained into the meaning of polynomials in relation to the number systems used. A remarkable fact appears: polynomials are very like integers. The parallel is almost exact. In the end, a fundamental theorem of algebra emerges to match that of arithmetic.

The difficulty is to define a polynomial without begging any questions. Reverse the order of the terms (for the moment) and write $a+bx+cx^2+\dots$. What is x? In view of what we have said, we hedge: leave x *undefined*. We then concentrate on the coefficients and say that a polynomial is just a *set of coefficients* $(a, b, c, \dots)$. The order of the coefficients matters; it must not be disturbed. We want, for example, (a, b, c) to mean $a+bx+cx^2$ but (a, c, b) to mean $a+cx+bx^2$. We have not said how many coefficients there are. However, we must agree to take only a finite number of non-zero coefficients. We can agree, further, to ignore any zero coefficients which follow the last non-zero one. It does not matter whether we write them or not, except to make clear which is the last non-zero coefficient.

The replacement set of the coefficients can be any number system. To have something specific to talk about, take the replacement set as the system of rationals. Hence, for illustration, we consider polynomials with rational coefficients.

Our *definition* of a polynomial is an ordered set of rational coefficients containing only a finite number of non-zero entries. *Rules* for sums and products are to be laid down. The rule for sums is easy:

$$(a_1, b_1, c_1, \dots)+(a_2, b_2, c_2, \dots)=(a_1+a_2, b_1+b_2, c_1+c_2, \dots) \quad\dots\dots(1)$$

The rule for products is not so obvious. For sets of three coefficients (the quadratic case), we can write:

$$(a_1, b_1, c_1) \times (a_2, b_2, c_2)$$
$$= (a_1a_2,\ a_1b_2 + a_2b_1,\ a_1c_2 + b_1b_2 + a_2c_1,\ b_1c_2 + b_2c_1,\ c_1c_2) \ldots (2)$$

and other such rules can be written for other cases. It all seems odd and arbitrary, but we are, in fact, just looking ahead.

To bring in 'x', we take as our guide the use of '$+i$' as a *place-marker* in a complex number, except that we plan to use several place-markers: '$+x$', '$+x^2$', '$+x^3$', ... As a notation, write

$$(a, b, c, \ldots) = a(+x)b(+x^2)c \ldots = a + bx + cx^2 + \ldots .$$

The rules (1) and (2) for sums and products begin to look familiar:

$$(a_1 + b_1x + c_1x^2 + \ldots) + (a_2 + b_2x + c_2x^2 + \ldots)$$
$$= (a_1 + a_2) + (b_1 + b_2)x + (c_1 + c_2)x^2 + \ldots$$

$$(a_1 + b_1x + c_1x^2) \times (a_2 + b_2x + c_2x^2)$$
$$= a_1a_2 + (a_1b_2 + a_2b_1)x + (a_1c_2 + b_1b_2 + a_2c_1)x^2 + (b_1c_2 + b_2c_1)x^3 + c_1c_2x^4.$$

Hence, the rules are seen to correspond to the ordinary operations of algebra; the use of place-markers $(+x,\ +x^2, \ldots)$, with x undefined, is vindicated. It is then found that the set of all polynomials obeys all the operational rules for sums (and differences) and all those for products, except that there are no reciprocals (and no division). In the set of polynomials, we add, subtract and multiply according to familiar rules; we are not, as yet, interested in dividing one polynomial by another. Polynomials form an *integral domain* with the same properties as the set of integers. A formal development is given in 15.2.

In taking stock, we can make a series of observations:

(i) There can be many zero coefficients in particular polynomials. For example:

$(0, 0, \ldots) = 0 + 0x + \ldots = 0$	the zero of the set of polynomials
$(1, 0, \ldots) = 1 + 0x + \ldots = 1$	the unity of the set of polynomials.

Further, $(a, 0, \ldots) = a$ (any rational); the set of polynomials includes all rationals.

(ii) The undefined $x, x^2, x^3, \ldots$ are to be interpreted:

$$(0, 1, 0, \ldots) = 0 + 1x + 0x^2 + \ldots = x$$
$$(0, 0, 1, 0, \ldots) = 0 + 0x + 1x^2 + 0x^3 + \ldots = x^2$$
$$(0, 0, 0, 1, 0, \ldots) = 0 + 0x + 0x^2 + 1x^3 + 0x^4 + \ldots = x^3 \ldots$$

By the product rule (2): $(0, 1, 0) \times (0, 1, 0) = (0, 0, 1, 0, 0)$ which translates into $x \times x = x^2$. This and similar results establish $x^2, x^3, \ldots$ as powers of x. Hence, the undefined x and all its integral powers are themselves polynomials.

(iii) The place-marker notation is fully justified. Not only are $x, x^2, x^3, \ldots$ polynomials in themselves, but the + sign means addition. We can operate with polynomials on familiar lines. For example:

$$\begin{aligned}(1-x+2x^2)(1+x+2x^2) &= 1+x+2x^2 \\ &\quad\; -x-\;x^2-2x^3 \\ &\quad\quad\; +2x^2+2x^3+4x^4 \\ \hline &= 1 \quad\; +3x^2 \quad\quad +4x^4.\end{aligned}$$

(iv) In our illustrative case, the coefficients are rational numbers. More strictly, they are a finite sequence of rationals and, if the last non-zero item is the coefficient of x^n, then n is an integer $(n \geqslant 0)$ called the *degree* of the polynomial. So:

$$f(x) = a + bx + cx^2 + \ldots + hx^{n-1} + kx^n \quad (k \neq 0)$$

is the general polynomial of degree n. Here n is a given integer $(n \geqslant 0)$ and $a, b, c, \ldots k$ are rationals. A polynomial of *zero degree* is a rational number a and the set of all polynomials includes the field of rationals. We may sometimes require a polynomial to be of *positive degree* $(n > 0)$.

(v) A polynomial is not itself a rational number, or indeed a number at all. In writing $f(x) = a + bx + cx^2 + \ldots$, we specify a set of coefficients $(a, b, c, \ldots)$ and leave x undefined. It is to be distinguished from:

$$f(\alpha) = a + b\alpha + c\alpha^2 + \ldots \quad \text{for a given rational } \alpha.$$

This *is* a rational number; $f(\alpha)$ is obtained from $\alpha, a, b, c, \ldots$ by algebraic processes. Further, $f(\alpha)$ can be extended to a real (or complex) number by simply making α a real (or complex) number. In the polynomial $f(x)$, x is undefined; in the value $f(\alpha)$, α is some specified rational, real or complex number.

To get a simpler notation, reverse again the order of the terms and write coefficients with subscripts. The general polynomial of degree n is:

$$f(x) = f_n x^n + f_{n-1} x^{n-1} + \ldots + f_1 x + f_0 \quad (n \geqslant 0, f_n \neq 0) \ldots\ldots\ldots (3)$$

where the f's are rationals. The leading coefficient can be factored out:

$$f(x) = f_n\left(x^n + \frac{f_{n-1}}{f_n}x^{n-1} + \dots + \frac{f_1}{f_n}x + \frac{f_1}{f_n}\right)$$

which is a case of the process known as *scalar multiplication* (see 15.2, and 3.9 Ex. 7). If we ignore the factor f_n, we can write (3) in alternative form:

$$f(x) = x^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0 \quad (n \geqslant 0) \quad \dots\dots\dots\dots(4)$$

where the a's are still rationals. The set of all polynomials, given by (3) or (4), is an integral domain, subject to all the operational rules with the sole exception that reciprocals are lacking.

The limitation of polynomials to those with rational coefficients is adopted here for purposes of illustration. More generally, the coefficients can be from any field F we care to specify:

DEFINITION: *Given a field F, a set of ordered coefficients from F containing a finite number of non-zero elements is a* **polynomial over the field F** *denoted $f(x) = f_nx^n + f_{n-1}x^{n-1} + \dots + f_1x + f_0$ where x is undefined, where f_r $(r = 0, 1, 2, \dots n)$ is an element of F and where $n \geqslant 0$.*
The set of polynomials over F is an integral domain, denoted $F[x]$. Usually F is the field of rational, real or complex numbers; but there are other possibilities as illustrated in 3.9 Ex. 11.

3.4. Rational fractions. The set $F[x]$ of polynomials $f(x)$ over a field F is not the end of this line of development. To complete, we have the analogy of the set of integers. The integral domain J is made into the field R of rationals by the process of forming the 'quotient field' (2.6). Ordered pairs (p, q) of integers are written, sums and products defined, and the pairs denoted p/q $(q \neq 0)$. The exposition can be repeated word for word, substituting 'polynomials' for 'integers'. Take two polynomials $f(x)$ and $g(x)$, write the ordered pair $\{f(x), g(x)\}$, define appropriate sums and products and identify the pair as the *rational fraction:*

$$\frac{f(x)}{g(x)} = \frac{f_nx^n + f_{n-1}x^{n-1} + \dots + f_1x + f_0}{g_mx^m + g_{m-1}x^{m-1} + \dots + g_1x + g_0} \text{ where } g(x) \neq 0.$$

The set of all such rational fractions is then found to satisfy all the operational rules (see 15.2). It is a field:

Definition: *Given the integral domain $F[x]$ of polynomials over a field F, a* **rational fraction** *is the ratio of two polynomials of $F[x]$, i.e. $f(x)/g(x)$ where $g(x) \neq 0$. The set of all rational fractions is a field, denoted $F(x)$.*

The field $F(x)$ includes the integral domain $F[x]$, the special cases where $g(x)=1$; it also includes F itself, the special cases where $f(x)=f_0$, $g(x)=1$.

Another way of looking at the set $F(x)$ is as the field obtained by the adjunction of x (and hence of all its powers) to the number field F with which we start. Again x is undefined, any additional element. This is the general process of adjunction (see 15.2).

Consider a particular case, the adjunction of the element $x=i$ to the field of real numbers. The result is the field of rational fractions in i with real coefficients. Things are made simpler here by the fact that $x^2=i^2=-1$, and hence that $x^3=i^3=-i$, $x^4=i^4=1, \ldots$. The rational fraction reduces to the ratio of f_0+if_1 to g_0+ig_1, and this further reduces to the form $a+ib$ (by multiplying numerator and denominator by g_0-ig_1). Hence, the complex number $a+ib$ is obtained as a particular rational fraction, the adjunction of $x=i$ to the field of real numbers.

We may proceed with polynomials and their ratios according to the familiar algebraic processes. Consider the rational fraction $(x^2+2x-1)/(x^2+1)$. Divide:

$$\begin{array}{rl} x^2+1\ ) & x^2+2x-1\ (\ 1 \\ & x^2 \qquad +1 \\ \hline & \qquad 2x-2 \end{array}$$

giving quotient 1 and remainder $2(x-1)$. Hence:

$$\frac{x^2+2x-1}{x^2+1} = 1 + 2\,\frac{x-1}{x^2+1}$$

by a process similar to that of reducing a rational to its lowest terms.

3.5. Polynomial functions. In the polynomial $f(x)$ of the set $F[x]$, we have taken x as undefined. We know only that x and all its powers are themselves polynomials of $F[x]$. We are entitled, however, to substitute anything we like for x and to follow through the calculations indicated by $f(x)=f_nx^n+f_{n-1}x^{n-1}+\dots+f_1x+f_0$. The results so obtained are now to be investigated.

Suppose that $f(x)=f_nx^n+f_{n-1}x^{n-1}+\dots+f_1x+f_0$ is a polynomial over the field of rationals. This does not mean that $f(x)$ is itself a rational. It is not *any* kind of number. It is, by definition, simply a sequence of (rational) coefficients. Suppose, however, that we substitute for x a rational α. Instead of $f(x)$, we have

$$f(\alpha)=f_n\alpha^n+f_{n-1}\alpha^{n-1}+\dots+f_1\alpha+f_0.$$

The *polynomial* $f(x)$ goes; in its place, we have $f(\alpha)$ which is a *rational number*. $f(\alpha)$ is obtained by arithmetical processes operating on rationals, α and all the f's. It is essential to keep the two things separate: $f(x)$ as a polynomial with rational coefficients and undefined x; $f(\alpha)$ a rational number obtained from a given rational number α. The same rational coefficients appear in $f(x)$ and $f(\alpha)$; this is the only link. So, if:

$$f(x)=x^2-\tfrac{5}{2}x+1$$

then: $f(\tfrac{1}{2})=(\tfrac{1}{2})^2-\tfrac{5}{2}(\tfrac{1}{2})+1=0$ and $f(1)=1^2-\tfrac{5}{2}1+1=-\tfrac{1}{2}$.

Similarly: $f(\tfrac{3}{2})=-\tfrac{1}{2}; f(2)=0; f(\tfrac{5}{2})=1; f(3)=\tfrac{5}{2}; \dots$

It is clear, in terms of elementary algebra, what we are doing here. We are making the calculations necessary for plotting the 'function':

$$y=x^2-\tfrac{5}{2}x+1.$$

The process of writing and graphing polynomial functions is, in fact, justified on the following lines. $f(x)$ is a given polynomial. Replace x by any rational α from the field R. The rational value $f(\alpha)$ is obtained; write it β:

$\beta=f(\alpha)$ a function of α over the field R of rationals.

More usually, α and β are written as x and y. This is in order, as long as we remember that x and y are rationals, that x is no longer the undefined x of a polynomial. Hence the *polynomial function:*

$$y=f(x)=f_nx^n+f_{n-1}x^{n-1}+\dots+f_1x+f_0$$

for x in the domain of all rationals.

With this achieved, another step is easily made. Replace x in the polynomial $f(x)$ by a *real number* α. Then

$$f(\alpha)=f_n\alpha^n+f_{n-1}\alpha^{n-1}+\dots+f_1\alpha+f_0$$

is also a real number, in the field R^*. Hence, the same polynomial function $y=f(x)$ can be written, except that it is defined for x in the domain of all real numbers. In drawing a graph of the function, we actually work with x as a rational. But, when we draw a smooth curve through the plotted points, we are implicitly taking x as a real number (see 1.4 above).

In the same way, the *rational fraction function:*

$$y = \frac{f(x)}{g(x)} = \frac{f_n x^n + f_{n-1}x^{n-1} + \ldots + f_1 x + f_0}{g_m x^m + g_{m-1}x^{m-1} + \ldots + g_1 x + g_0} \quad (g_m \neq 0)$$

is defined over the domain either of all rationals or of all real numbers. For example:

(i) $y=(x^2+2x-1)/(x^2+1)$ for x in the domain of all real numbers. The graph (Fig. 3.5) can be plotted from a selection of rational values of x and the corresponding values of y:

x	-3	-2	-1	$-\frac{1}{2}$	0	1	2	$\frac{5}{2}$	3
y	$\frac{1}{5}$	$-\frac{1}{5}$	-1	$-\frac{7}{5}$	-1	1	$\frac{7}{5}$	$\frac{41}{29}$	$\frac{7}{5}$

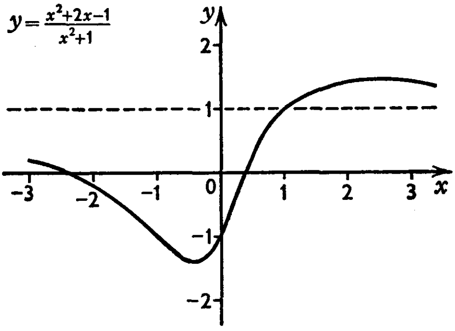


FIG. 3.5

As will be seen later (Chapter 9), the 'limit' of y is 1, as x increases indefinitely in each direction.

(ii) $y=(x^3-1)/(x-1)$ for x in the domain of all real numbers $(x \neq 1)$. Here, division of $x-1$ into x^3-1 gives $(x^3-1)/(x-1)=x^2+x+1$. The polynomial x^2+x+1 is defined for all real x and the rational fraction $(x^3-1)/(x-1)$ for all real x *except* $x=1$; the two are equivalent *except* at $x=1$. (See 3.9 Ex. 9.)

The concept of a *function* will be introduced in a wider context in Chapter 7 and, from Chapter 9 onwards, we shall be largely concerned with functions $f(x)$ of a real variable x, i.e. for x in the domain of all real numbers. We shall then have polynomials or rational fractions ready to hand as illustrations.

Meanwhile, one further extension can be made; it suggests itself in the present context. In the polynomial $f(x)$ or the rational fraction $f(x)/g(x)$, replace x by a *complex number* α. Then

$$f(\alpha)=f_n\alpha^n+f_{n-1}\alpha^{n-1}+\ldots+f_1\alpha+f_0$$

is itself a complex number, and so is $f(\alpha)/g(\alpha)$. We are thus lead to the concept of a *complex function*, i.e. the function $\beta=f(\alpha)$ or $f(\alpha)/g(\alpha)$ over the field C of complex numbers. The usual notation for a complex number is z. Hence we can write the polynomial function $f(z)$, or the rational fraction function $f(z)/g(z)$, for z in the domain of all complex numbers. The idea of a function of a complex variable is generally regarded as difficult or advanced. This is nonsense. The concept of $f(z)$ as a function of a complex variable is no more difficult than that of $f(x)$ as a function of a real variable. The difference lies in the field of numbers over which the functions are defined and hence in the varying techniques for handling them (see 3.9 Ex. 12).

This line of development will be taken up later. At the moment, we have more urgent business on hand.

3.6. Roots of polynomial equations. To pass from a polynomial $f(x)$ to a polynomial equation and its roots is not quite as simple as it appears. It is not just a matter of writing $f(x)=0$ and looking for $x=\alpha$ as a root. This is over-working x. Strictly, $f(x)=0$ is the zero polynomial, all coefficients zero; it is *not* an equation. However, the correct procedure is ready to hand, using the concepts of 3.5. A zero α of a given polynomial $f(x)$ is a number α such that the number $f(\alpha)$ is zero. Consider a polynomial of positive degree* defined over the field of rationals:

DEFINITION: *The polynomial $f(x)=x^n+a_{n-1}x^{n-1}+\ldots+a_1x+a_0$ with rational coefficients and of degree $n>0$ has a* **zero** *α if $f(\alpha)=0$.*
If we then care to say that the polynomial equation $f(x)=0$ has a

* A polynomial of zero degree $f(x)=f_0$ has no zeros; the question does not arise. For $f(x)$ of positive degree ($n>0$), we write the leading coefficient unity, as in (4) of 3.3, since the removal of a constant factor does not influence zeros.

root $x=\alpha$, this is in order, but we must remember that we are using short-hand. In a graphical representation (for real x) a zero of $y=f(x)$ occurs where the corresponding curve crosses Ox, and it gives a root of the equation $f(x)=0$.

The definition of a zero or root is completely neutral as to what kind of number a is. It can be from the field of rationals R, real numbers R^* or complex numbers C according to our choice. The function $y=f(x)$ can be defined over any one of these fields. All that we have specified is that $f(x)$ is a polynomial (of degree $n>0$) with rational coefficients.

For the moment suppose α is *rational*. The following results, much used in elementary algebra, are proved very simply:

REMAINDER THEOREM: *If the polynomial $f(x)$ of degree $n>0$, and with rational coefficients, is divided by $x-\alpha$, then the remainder is $f(\alpha)$, i.e. $f(x)=(x-\alpha)g(x)+f(\alpha)$ for some polynomial $g(x)$ of degree $n-1$.*

Proof: let R be the remainder so that $f(x)=(x-\alpha)g(x)+R$ for any rational x. Put $x=\alpha$ so that $f(\alpha)=0\times g(\alpha)+R$, i.e. $R=f(\alpha)$.

Q.E.D.

As a corollary, it follows that:

THEOREM: *The polynomial $f(x)$ of degree $n>0$, and with rational coefficients, has a zero α if and only if $x-\alpha$ divides $f(x)$, i.e. if and only if $f(x)=(x-\alpha)g(x)$ for some polynomial $g(x)$ of degree $n-1$.*

Proof: directly, if $x-\alpha$ divides $f(x)$, the remainder $R=f(\alpha)=0$ and α is a zero of $f(x)$. Conversely, if α is a zero of $f(x)$, then $f(\alpha)=0$ and $f(x)=(x-\alpha)g(x)$, i.e. $x-\alpha$ divides $f(x)$. Q.E.D.

Notice that this result merely writes $g(x)$ as some polynomial of degree $(n-1)$. Nothing is implied about the rational number $g(\alpha)$. If $g(\alpha)\neq 0$, then $(x-\alpha)$ is not a factor of $g(x)$ and $f(x)$ has only one factor $(x-\alpha)$. Here α is a *single root* of $f(x)=0$. If $g(\alpha)=0$, then $(x-\alpha)$ is a factor of $g(x)$ and $f(x)$ has two or more factors $(x-\alpha)$. Here α is a *multiple root* of $f(x)=0$. (See 3.9 Ex. 13.)

As long as we stick to α as a rational, there is no implication that $f(x)=0$ has a root at all, or (if it has one) that there are further roots arising from the quotient polynomial $g(x)$. Certainly, there is no implication that $f(x)=0$ has n rational roots. We can, however, get out of this situation.

Suppose now that α is *real*. We can look for a real zero of $f(x)$, i.e.

a real α so that the real number $f(\alpha)=0$. The theorems above still hold, and the proofs are formally unchanged. There is, however, a subtle difference. If α is a real zero of $f(x)$, then $(x-\alpha)$ divides $f(x)$:

$$f(x)=(x-\alpha)g(x)$$

where $g(x)$ is a 'polynomial' of degree $(n-1)$. The difference is that, whereas, $f(x)$ has only rational coefficients, $(x-\alpha)$ and hence $g(x)$ have coefficients from the wider field of real numbers. In short, $g(x)$ is *not* a polynomial from the set $F[x]$ of polynomials with rational coefficients. Irrational coefficients like $\surd 2$ can appear both in $(x-\alpha)$ and in $g(x)$. Indeed, if α is irrational, then $g(x)$ *must* contain irrational coefficients.

It is clear, from what we know about the quadratic, that even real values of α do not exhaust the possibilities. We look further, for α as a *complex* zero of $f(x)$, such that the complex number $f(\alpha)=0$. Again the theorems above still hold. In factorising $f(x)$ into $(x-\alpha)g(x)$, it is possible for $g(x)$ to have complex coefficients; indeed if $\alpha=a+ib$ $(b\neq 0)$, then $g(x)$ *must* contain such coefficients.

What remains to be shown is that we need go no further, i.e. that all the roots of $f(x)=0$ are in the field of complex numbers and that there are precisely n of them. This is the fundamental theorem of algebra, concerned (as is the corresponding fundamental theorem of arithmetic) with factorisation. Various proofs of the theorem are available but none of them is easy; indeed it seems not possible to provide a proof purely in algebraic terms. It is remarkable that algebra would appear to rest on a non-algebraic basis.

3.7. The fundamental theorem of algebra. The analogy between the integral domain $F[x]$ of polynomials $f(x)$ and the integral domain J of integers is our guide. We start off, exactly as in 3.1 for integers, with the idea of getting the highest common factor (H.C.F.) of two polynomials, of getting a division algorithm, and of factorising a polynomial into prime (or irreducible) factors. The factors then lead to the roots of $f(x)=0$.

Take $f(x)=x^n+a_{n-1}x^{n-1}+\ldots+a_1x+a_0$ as a polynomial of degree $n>0$ and with rational coefficients and $g(x)$ as a similar polynomial but of degree $m>0$. The H.C.F. of $f(x)$ and $g(x)$ is the polynomial of highest degree which divides both. A division algorithm is developed to isolate it. By the same argument as in 3.1, take $m<n$ and divide $g(x)$ into $f(x)$:

$$f(x)=q_1(x)g(x)+r_1(x)$$

where $q_1(x)$ is the quotient polynomial and $r_1(x)$ of degree $<m$ is the remainder. As long as the remainder is of positive degree, we carry on dividing: $r_1(x)$ into $g(x)$, then the new remainder into $r_1(x)$, and so on. Since the successive remainders are of decreasing degree, it must happen sooner or later that a remainder of zero degree results. The last remainder (before this stage) is the H.C.F. Two examples illustrate:

(i) $$x^2+2x+1 \quad \text{and} \quad x^2-1.$$

$$\left.\begin{array}{l} \text{First } \dfrac{x^2+2x+1}{x^2-1} \text{ has quotient 1 and remainder } 2(x+1); \\[2ex] \text{then } \dfrac{x^2-1}{x+1} \text{ has quotient } (x-1) \text{ and no remainder} \end{array}\right\} \begin{array}{c}\text{H.C.F.}\\ =(x+1).\end{array}$$

Or: $(x^2+2x+1)=(x^2-1)+2(x+1)$ and $(x^2-1)=(x-1)(x+1)$.

(ii) $$x^4+3x^2+2 \quad \text{and} \quad x^3-x^2+2x-2.$$

$$\left.\begin{array}{l} \text{First } \dfrac{x^4+3x^2+2}{x^3-x^2+2x-2} \text{ has quotient } (x+1) \\ \qquad \text{and remainder } 2(x^2+2); \\[1ex] \text{then } \dfrac{x^3-x^2+2x-2}{x^2+2} \text{ has quotient } (x-1) \\ \qquad \text{and no remainder} \end{array}\right\} \quad \text{H.C.F.}=(x^2+2).$$

Or: $$(x^4+3x^2+2)=(x+1)(x^3-x^2+2x-2)+2(x^2+2)$$

and $$(x^3-x^2+2x-2)=(x-1)(x^2+2).$$

Jobbing backward, as in 3.1, the H.C.F. is expressed as:

$$\text{(polynomial) } f(x)+\text{(polynomial) } g(x)$$

proving the result:

THEOREM: *If $c(x)$ is the H.C.F. of two polynomials $f(x)$ and $g(x)$ of positive degree, then there exist polynomials $\phi(x)$ and $\psi(x)$ so that*

$$\phi(x)f(x)+\psi(x)g(x)=c(x) \quad \text{.........................(1)}$$

In example (ii) above:

$$x^4+3x^2+2=(x+1)(x^3-x^2+2x-2)+2(x^2+2)$$

i.e. $\frac{1}{2}(x^4+3x^2+2)-\frac{1}{2}(x+1)(x^3-x^2+2x-2)=x^2+2$ (H.C.F.)

so that $\phi(x)=\frac{1}{2}$ and $\psi(x)=-\frac{1}{2}(x+1)$.

A polynomial $p(x)$ is *irreducible* in the field R of rationals if $p(x)$ is of degree $n>0$ and if it has no polynomial factors with rational coefficients other than 1 and $p(x)$ itself. Exactly as for integers in 3.1, it follows from (1) that:

THEOREM: *Every polynomial $f(x)$ of positive degree and with rational coefficients can be uniquely factorised into a product of irreducible factors:*

$$f(x) = p_1(x)p_2(x) \dots p_i(x) \dots\dots\dots\dots (2)$$

We have followed the argument used for integers; but the result we have obtained is not the end of the story as it was for integers. One kind of irreducible polynomial is $(x-\alpha)$, for α rational. Among the factors of (2), there may be some of linear type $(x-\alpha)$. It is, however, not necessary that all of the factors, or indeed any of them, should be of this form. Quadratics like (x^2-2) and (x^2+1) are irreducible *in the field R of rationals*. It is necessary, therefore, to extend the range of factors so that their coefficients are from the field of complex numbers. Then a polynomial $p(x)$ irreducible in the field of rationals may become reducible (into factors) in the field of complex numbers,

$$\text{e.g.}\quad x^2-2=(x-\surd 2)(x+\surd 2) \quad\text{and}\quad x^2+1=(x-i)(x+i).$$

At the same time, the limitation that the polynomial $f(x)$ has rational coefficients can be relaxed. The fundamental theorem of algebra states that, if a polynomial with rational, real or complex coefficients is considered over the field of complex numbers, then the only irreducible factors are of linear type $(x-\alpha)$, for α complex. In its simplest form:

FUNDAMENTAL THEOREM OF ALGEBRA: *Every polynomial $f(x)$ of positive degree and with rational, real or complex coefficients is such that $f(\alpha)=0$ for some complex number α.*

This means that there is a complex root α of $f(x)=0$ and hence, from 3.6 above, that $(x-\alpha)$ is a factor of $f(x)$ for some complex α. Notice that complex roots automatically include (as special cases) all real roots, both rational and irrational.

The proof of the theorem (apparently) cannot be given in algebraic terms. Those available involve topological concepts; one is sketched in 15.2.

The fundamental theorem can be developed into a form more directly usable in practice. The first step is to write the theorem in the form: $f(x)=0$ of degree $n(>0)$ has a complex root α_1 and so a factor $(x-\alpha_1)$, giving $f(x)=(x-\alpha_1)g(x)$. If $n=1$, then $g(x)=1$ and the factorisation is complete: $f(x)=x-\alpha_1$. If $n>1$, then $g(x)$ is a polynomial of degree $(n-1)$, with rational, real or complex coefficients. Hence, $g(x)=0$ of degree $(n-1)>0$ has a complex root α_2 and so a factor $(x-\alpha_2)$, giving $f(x)=(x-\alpha_1)(x-\alpha_2)h(x)$. Here $h(x)=1$ if $n=2$ and otherwise $h(x)$ is a polynomial of degree $(n-2)>0$. This process continues until a residual polynomial is obtained equal to unity; this happens after n steps and so:

$$f(x)=(x-\alpha_1)(x-\alpha_2)\;\ldots\;(x-\alpha_n).$$

Hence, $f(x)$ of degree n has *exactly* n factors and $f(x)=0$ has *exactly* n roots in the field of complex numbers.

One further step can be taken when $f(x)$ has rational or real coefficients. If $\alpha=a+ib$ is a complex number, write $\alpha^*=a-ib$ and call it the *conjugate* of α. Then

$$\alpha+\alpha^*=(a+ib)+(a-ib)=2a$$

and $$\alpha\times\alpha^*=(a+ib)(a-ib)=a^2+b^2$$

which are real. On an Argand Diagram (Fig. 3.7) if α is represented by P, then α^* is P^*, the reflection of P in Ox; the sum of α and α^* is the point Q on Ox. Since OP and OP^* make equal and opposite angles with Ox, the product of α and α^*(adding the angles) is also a point on Ox.

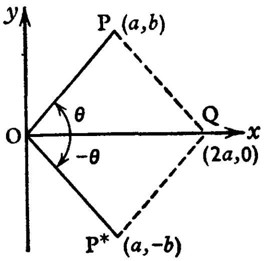


FIG. 3.7

In the relation obtained for $f(x)$ with rational or real coefficients:

$$f(x)=x^n+a_{n-1}x^{n-1}+\ldots+a_1x+a_0=(x-\alpha_1)(x-\alpha_2)\ldots(x-\alpha_n)$$

replace each number (coefficient) by its conjugate:

$$x^n+a_{n-1}{}^*x^{n-1}+\ldots+a_1{}^*x+a_0{}^*=(x-\alpha_1{}^*)(x-\alpha_2{}^*)\ldots(x-\alpha_n{}^*).$$

Any real coefficient is equal to its conjugate ($\alpha=\alpha^*$ if $b=0$). This is so for all the coefficients on the left (and for such of the α's on the right which are real). Hence:

$$(x-\alpha_1)(x-\alpha_2)\ldots(x-\alpha_n)=(x-\alpha_1{}^*)(x-\alpha_2{}^*)\ldots(x-\alpha_n{}^*)$$

i.e. the set α_1, α_2, ... α_n is simply a shuffling of the set $\alpha_1{}^*$, $\alpha_2{}^*$, ... $\alpha_n{}^*$, real values remaining in the same place but complex values being shifted. Hence, if complex α is a root of $f(x)$, then so is the conjugate α^*.

The final and practical result obtained is:

> Every polynomial $f(x)$ of degree $n>0$, and with rational or real coefficients, has exactly n linear factors in the field of complex numbers:
>
> $$f(x)=(x-\alpha_1)(x-\alpha_2)\ldots(x-\alpha_n)$$
>
> and $f(x)=0$ has exactly n roots $\alpha_1, \alpha_2, \ldots \alpha_n$. Some or all of the roots may be real. Other roots occur in pairs of conjugate complex numbers. If the degree n is odd, at least one and generally an odd number of roots are real. If the degree n is even, there may be no real roots and, in general, an even number of real roots.

As a final point, the factor $(x-\alpha)$ may appear only once, in which case we have a single root. It may, however, appear twice or more often, corresponding to a multiple root. Some examples illustrate:

(i) $x^3-\frac{5}{2}x^2+\frac{1}{2}=(x-\frac{1}{2})(x^2-2x-1)$ irreducible in the field of rationals

$$=(x-\tfrac{1}{2})(x-1-\surd 2)(x-1+\surd 2)$$

i.e. $x^3-\frac{5}{2}x^2+\frac{1}{2}=0$ has three real roots $\frac{1}{2}$, $1\pm\surd 2$, one rational and two irrational.

(ii) $x^3+x+10=(x+2)(x^2-2x+5)$ irreducible in the field of real numbers

$$=(x+2)(x-1-2i)(x-1+2i)$$

i.e. $x^3+x+10=0$ has one real root -2 and the conjugate complex pair $(1\pm 2i)$.

(iii) $x^5-x^4-2x^3+2x^2+x-1=(x-1)(x-1)(x-1)(x+1)(x+1)$

$$=(x-1)^3(x+1)^2.$$

The corresponding equation has all five roots real, one triple (1) and one double (-1).

(iv) $x^4-4x^3+10x^2-12x+9=(x^2-2x+3)^2$ irreducible for real numbers

$$=(x-1-i\surd 2)^2(x-1+i\surd 2)^2.$$

The equation has four complex roots, a double conjugate complex pair $(1\pm i\surd 2)$.

3.8. The nth roots of unity. It may seem rather like gilding the lily to ask for the nth roots of unity, i.e. the zeros of the polynomial $x^n - 1$ or the roots of the equation $x^n = 1$. Surely, it may be said, the root 1 (the zero $x = 1$) is enough. However, a question emerges immediately. By the result just obtained, the polynomial $x^n - 1$ of degree n has exactly n zeros. One of them is 1; what are the other $(n - 1)$? The investigation of the subject has a fascination for pure mathematicians, and to pursue it far would take us deep into number theory. On the other hand, a quick look at the problem is not a useless exercise; it provides an illustration of a 'transformation group' (6.4 below).

Let us attempt a direct attack. We require n roots ω such that $\omega^n = 1$. One root is easy: $\omega = 1$. Take out the factor $(x - 1)$ from $(x^n - 1)$:

$$\frac{x^n - 1}{x - 1} = x^{n-1} + x^{n-2} + \ldots + x + 1$$

which is just a geometric progression (in reverse order) with common ratio x and sum of n terms $x^n - 1/x - 1$ as shown. Hence the other $(n - 1)$ roots come from:

$$x^{n-1} + x^{n-2} + \ldots + x + 1 = 0.$$

This looks easy, but isn't. If n is odd, nothing else is obvious enough to try. If n is even $(n = 2m)$, then $\omega = -1$ is a root as well as $\omega = 1$:

$$\frac{x^{2m-2} + x^{2m-2} + \ldots + x + 1}{x + 1} = x^{2m-2} + x^{2m-4} + \ldots + x^2 + 1.$$

Or, from the beginning:

$$(x^{2m} - 1)/(x^2 - 1) = x^{2m-2} + x^{2m-4} + \ldots + x^2 + 1$$

i.e. a geometric progression of m terms with common ratio x^2. But we are back with essentially the same polynomial for the other roots, in x^2 now:

$$(x^2)^{m-1} + (x^2)^{m-2} + \ldots + (x^2) + 1.$$

Again there is nothing obvious to try.

As an alternative, let us try to sneak up on the problem, by working out the solution for small values of $n \geqslant 1$:

$\underline{n = 1}$:	$x - 1$	One root, $\omega = 1$
$\underline{n = 2}$:	$x^2 - 1 = (x - 1)(x + 1)$	Two roots, $\omega = \pm 1$

$\underline{n=3}$: $x^3-1=(x-1)(x^2+x+1)$ Three roots, $\omega=1, \frac{1}{2}(-1\pm i\sqrt{3})$ where the pair of conjugate complex roots are from the quadratic

$\underline{n=4}$: $x^4-1=(x^2-1)(x^2+1)$ Four roots, $\omega=\pm 1, \pm i$.

For $n>4$, we cannot continue, in purely algebraic terms.* But we can make progress in graphical terms with the aid of the Argand Diagram for complex numbers, here the nth roots of unity. We get, first, a broad hint from the following facts about the roots where $n=2$, 3 and 4. For $n=2$, the two roots (± 1) can be shown: $\omega=-1$, $\omega^2=1$. For $n=3$, the three roots are:

$$\omega=\tfrac{1}{2}(-1+i\sqrt{3}),\ \omega^2=\tfrac{1}{4}(-1+i\sqrt{3})^2=\tfrac{1}{2}(-1-i\sqrt{3}),\ \omega^3=1.$$

For $n=4$, the four roots are: $\omega=i$, $\omega^2=-1$, $\omega^3=-i$, $\omega^4=1$. This suggests (no more) that, if ω is one of the roots of $x^n=1$ not previously obtained as a lower root of unity, then the n roots are: ω, ω^2, ω^3, ... ω^n. By definition, $\omega^n=1$ so that the last root is 1, as required.

To indicate (if not to prove strictly) that this is so, we use an Argand Diagram. The roots of $x^2=1$ are $\omega=-1$ at A' and $\omega^2=1$ at A. The roots of $x^3=1$ are:

$$\omega=\frac{1}{2}(-1+i\sqrt{3}) \quad \text{at } C\left(-\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$$

$$\omega^2=\frac{1}{2}(-1-i\sqrt{3}) \quad \text{at } D\left(-\frac{1}{2}, -\frac{\sqrt{3}}{2}\right)$$

and $\omega^3=1$ at A.

To locate C and D (Fig. 3.8a), note first that D is the reflection of C in Ox (being conjugate). Let CD cut Ox in N. The triangle ONC is right-angled, with $ON=\dfrac{1}{2}$, $NC=\dfrac{\sqrt{3}}{2}$ and $OC=1$. The angle NOC is 60° ($\cos 60°=ON/OC=\frac{1}{2}$); and angle AOC is 120°. Hence C is 120° round the unit circle from A, and D is another 120° round from C. The three cube roots of unity are:

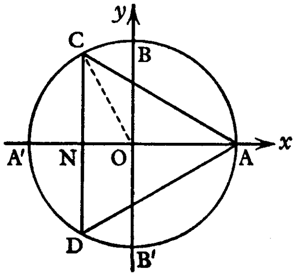


FIG. 3.8a

* Indeed, it is a general point that polynomial equations of degree 5 and higher are of a different nature from those of degree 4 and lower; there seems to be a 'sound barrier' at $n=5$.

$C(\omega)$, $D(\omega^2)$ and $A(\omega^3)$, making up an equilateral triangle inscribed in the unit circle. Finally, the roots of $x^4=1$ are:

$$\omega=i \text{ at } B;\ \omega^2=-1 \text{ at } A';\ \omega^3=-i \text{ at } B';\ \text{and } \omega^4=1 \text{ at } A.$$

They form a square in the unit circle: $B(\omega)$, $A'(\omega^2)$, $B'(\omega^3)$ and $A(\omega^4)$.

The generalisation is clear: the nth roots of unity correspond to the vertices of a regular n-sided polygon inscribed in the unit circle, the last vertex being at A. The polygon is $CDEF \ldots A$ in Fig. 3.8b, where the $\angle AOC=\frac{1}{n}360°$. To indicate that this is so: let C correspond to a complex number ω. Then $\omega^2=\omega\times\omega$ is the point P with $OP=OC^2=1$ and $\angle AOP=2\angle AOC=\frac{2}{n}\ 360°$ (see 2.5). The point P is D. And so on round the unit circle, until $\omega^n=1$ is obtained as the point A. Hence, the complex number ω, at the point C, is one of the nth roots of unity. So are all the powers of ω up to ω^n. For:

$$(\omega^2)^n=(\omega^n)^2=1^2=1;\ \ldots$$

The general result is:

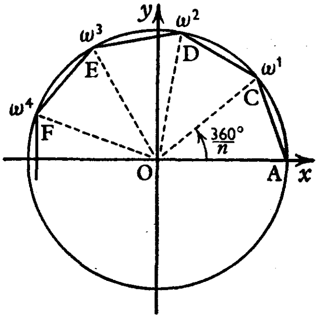


FIG. 3.8b

THEOREM: *The nth roots of unity, complex numbers ω such that $\omega^n=1$, are shown on an Argand Diagram as the vertices of a regular n-sided polygon $CDEF \ldots A$ inscribed in the unit circle. They can be expressed:*

$$\omega,\ \omega^2,\ \omega^3,\ \ldots\ \omega^n=1$$

where ω is the complex number of the first vertex C with $\angle AOC=360°/n$.

For $n>4$, it is not easy to represent ω and its powers algebraically in terms of (e.g.) surds such as square roots. It is still true that $\omega=a+ib$ for some real a and b and rational approximations to a and b are given by trigonometric tables.* However, the general result in diagrammatic form is often enough.

3.9. Exercises

1. Express 1092 and 330 as products of primes and show that the H.C.F. is 6. Check by the process of repeated division (Division Algorithm) and show that: $13\times1092-43\times330=6$.

* Since $a=\cos(360°/n)$ and $b=\sin(360°/n)$.

2. *Gaussian integers.* The set $\alpha = m + in$ (m and n integers) satisfies all the operational rules of 2.2 except that reciprocals are lacking, i.e. it is an integral domain. Check the rules for addition, noting that zero is $0 = 0 + i \times 0$ and that $-\alpha = (-m) + i(-n)$. Check the first three rules for products, note that unity is $1 = 1 + i \times 0$, but that this does not provide a reciprocal α^{-1} of α. Check the distributive rule. Finally show that there are no zero divisors by writing $\alpha = m + in$ and $\beta = p + iq$ and showing that $\alpha\beta = 0$ implies either

$$m = n = 0 \ (\alpha = 0) \quad \text{or} \quad p = q = 0 \ (\beta = 0)$$

or both. (Equate real and imaginary parts.)

3. Show that $13 = (2+3i)(2-3i) = (3+2i)(3-2i)$ and two other pairs of factors obtained by multiplying those written by -1. Deduce that 13 is not prime as a Gaussian integer, being uniquely factored: $13 = (2+3i)(2-3i)$, allowing for the four units ± 1, $\pm i$.

4. Show that $1 - 3i = (1-i)(2-i)$ and $2i = (1+i)^2$ are unique factors.

5. *Polynomials with integral coefficients.* Illustrate the limitations of a polynomial with coefficients confined to integers by reference to

$$2x^2 - x - 3 = 2(x^2 - \tfrac{1}{2}x - \tfrac{3}{2}) \quad \text{and} \quad x^2 - \tfrac{1}{2}x - \tfrac{3}{2} = \tfrac{1}{2}(2x^2 - x - 3).$$

Generalise to show (i) that a polynomial with integral coefficients cannot generally be written with leading coefficient unity, but (ii) that a polynomial with rational coefficients can always be written (apart from a rational factor) as a polynomial with integral coefficients.

6. From (2) of 3.3 for products of polynomials, show that:

$$(0, 1, 0) \times (0, 1, 0) = (0, 0, 1, 0, 0); \quad (0, 1, 0) \times (0, 0, 1) = (0, 0, 0, 1, 0);$$
$$(0, 0, 1) \times (0, 0, 1) = (0, 0, 0, 0, 1)$$

and interpret as $x \times x = x^2$, $x \times x^2 = x^3$, $x^2 \times x^2 = x^4$.

7. *Scalar Multiplication.* From (2) of 3.3 show that

$$(k, 0, 0)\ (a, b, c) = (ka, kb, kc).$$

Interpret as $k(a + bx + cx^2) = (ka) + (kb)x + (kc)x^2$ and examine the case $k = \dfrac{1}{c}$. The polynomial $(k, 0, 0) = k$, a rational *scalar*. This result is 'scalar multiplication' of polynomials, a process of very wide application.

8. Show: $\dfrac{1+x^2}{1-x^2} = \dfrac{2}{1-x^2} - 1$; $\dfrac{1-x^3}{1-x^2} = x + \dfrac{1}{1+x}$; $\dfrac{1+x^6}{1-x^2} = \dfrac{2}{1-x^2} - (1 + x^2 + x^4)$.

Taking each rational fraction as a function of x, write an appropriate domain of x in each case. Is the domain different in any of the reduced forms above?

9. Plot the graph of $y = (x^3 - 1)/(x - 1)$ defined on the domain of all real $x (x \neq 1)$. Show that it is identical with the graph of the quadratic $x^2 + x + 1$ except that the point where $x = 1$ is missing. Deduce that the range of y is $y > \frac{3}{4}$ $(y \neq 3)$.

10. Show that the graph of $y = \dfrac{x^2 + 2x + 2}{x^2 - 1} = 1 + \dfrac{2x + 3}{x^2 - 1}$ is of form indicated in

Fig. 3.9. Why is y not defined at $x = \pm 1$? With the idea of a limit (Chapter 9), we say that $y \to 1$ as $x \to \pm\infty$, giving an 'asymptote' $y = 1$.

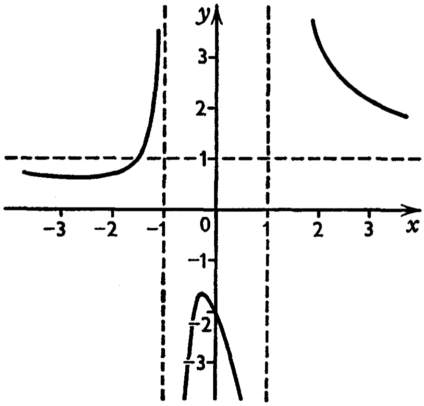


FIG. 3.9

*11. *Polynomials over the field of integers* (*mod n*). Generalise the results of 2.9 Ex. 22 and 23 to show that polynomials can be defined over the integers (mod n) but that they are only well-behaved if n is prime and the integers (mod n) form a field. What goes wrong if n is not prime? Consider the field of integers (mod 2) and show that there are then only four quadratics: x^2, x^2+1, x^2+x and x^2+x+1. Noting that $(x+1)^2 = x^2+1$ (mod 2), show that there is only one irreducible quadratic among the four.

12. *Polynomial function of a complex variable.* Consider $f(z) = z^2 - 1$ as a function of $z = x + iy$. Write $f(z)$ as the complex value $X + iY$ and show that $X = x^2 - y^2 - 1$ and $Y = 2xy$ (by equating real and imaginary parts). If $z^2 - 1 = 0$, show that $X = 0$, $Y = 0$ together, i.e. $x = \pm 1$, $y = 0$, i.e. $z = \pm 1$. Deduce that a polynomial equation in a *complex* variable can have *real* roots. Examine $f(z) = z^2 + 1$ similarly and show that $z^2 + 1 = 0$ has roots $z = \pm i$.

13. *Multiple roots.* Show that $f(x) = 0$, a polynomial equation with rational coefficients, has a double rational root α if and only if $f(x) = (x-\alpha)^2 g(x)$ where $g(x)$ is a polynomial with rational coefficients such that $g(\alpha) \neq 0$. Extend to an r-fold root α. How is the result affected if α is real or complex? Establish the following and find the other roots (if any) in each case.

Polynomial	*Multiple roots*	*Polynomial*	*Multiple roots*
$x^3 - 3x^2 + 4$	2 twice	$x^4 + 2x^3 - 2x - 1$	-1 three times
$x^4 - 4x^2 + 4$	$\begin{cases} \sqrt{2} \text{ twice} \\ -\sqrt{2} \text{ twice} \end{cases}$	$x^4 - 4x^3 + 8x^2 - 8x + 4$	$\begin{cases} 1+i \text{ twice} \\ 1-i \text{ twice} \end{cases}$

14. Show that $x-1$ is the H.C.F. of x^4-x^3+x-1 and x^3-x^2+x-1 by dividing out:

$$\begin{aligned} x^4-x^3+x-1 &= x(x^3-x^2+x-1)-(x^2-2x+1)\\ x^3-x^2+x-1 &= (x+1)(x^2-2x+1)+2(x-1)\\ x^2-2x+1 &= (x-1)(x-1) \quad \text{(with no remainder).}\end{aligned}$$

Hence show: $\phi(x)(x^4-x^3+x-1)+\psi(x)(x^3-x^2+x-1)=x-1$

where $\phi(x)=\frac{1}{2}(x+1)$ and $\psi(x)=-\frac{1}{2}(x^2+x-1)$.

15. Show that the following factors are irreducible in the field of rationals:

$$x^4-x^3+x-1=(x-1)(x+1)(x^2-x+1);\ x^3-x^2+x-1=(x-1)(x^2+1).$$

Write the factors of the quadratics in the field of complex numbers and hence find all zeros of each of the original polynomials.

16. If $x^n+a_{n-1}x^{n-1}+a_{n-2}x^{n-2}+\ldots+a_1x+a_0=0$ is any polynomial equation, establish the following criteria as practical guides:

(i) $x=0$ is a root if $a_0=0$

(ii) $x=1$ is a root if the coefficients sum to zero: $1+a_{n-1}+a_{n-2}+\ldots=0$

(iii) $x=-1$ is a root if the coefficients, alternating in sign, sum to zero:

$$1-a_{n-1}+a_{n-2}-a_{n-3}+\ldots=0.$$

Examine the polynomials of Ex. 13 and 15 above in the light of these criteria.

17. Show that a polynomial of degree $2n$ containing only even powers ($x^2, x^4, \ldots x^{2n}$) has zeros to be found from a polynomial of degree n in x^2. Illustrate by showing that x^4-4x^2+4 has zeros $x^2=2$ (twice), i.e. $x=\pm\sqrt{2}$ (each twice). What can be said of a polynomial of degree $2n+1$ containing no constant term and only odd powers ($x, x^3, \ldots x^{2n+1}$)? Show $x^5-2x^3+2x=0$ has roots 0, $\pm\sqrt{(1\pm i)}$.

*18. *Residue classes.* Divide the set J of all integers into five (non-overlapping and exhaustive) subsets J_r ($r=0, 1, 2, 3, 4$) according to the remainder r on division of an integer by 5. J_0 is $\{\ldots -10, -5, 0, 5, 10, \ldots\}$; write the other four. Show that the J_r's can be represented sufficiently by the corresponding r's and that the set of integers $\{0, 1, 2, 3, 4\}$ (mod 5) results. The J_r's are called *residue classes* of J. Carry out the same process in dividing the set $F[x]$ of all polynomials over the field of rationals into residue classes according to the remainder on division by x^2-2. Show that each residue class corresponds to a linear polynomial $ax+b$, for various rational a and b.

*19. *Field of polynomials* (*mod* x^2-2). Continuing from Ex. 18, show that, if $ax+b$ is the remainder on dividing $f(x)$ by x^2-2, then $ax+b$ is obtained by substituting $x^2=2$, $x^3=2x$, $x^4=4$, $x^5=4x$, ... in $f(x)$, a Remainder Theorem similar to that of 3.6. Hence show that the remainders obey all the operational rules for + and ×, including reciprocals and division. In particular:

$$(ax+b)(cx+d)=(ad+bc)x+(2ac+bd);\ \frac{ax+b}{cx+d}=\frac{(bc-ad)x+(2ac-bd)}{2c^2-d^2}.$$

(For the second, multiply numerator and denominator by $cx-d$; for both, put $x^2=2$.) Hence the set of remainders is a field, the field of polynomials (mod

$x^2 - 2$). The elements follow all elementary algebraic processes, provided x^2 is written 2.

*20. *Other polynomial fields.* As in Ex. 19, obtain the field of polynomials (mod x^2+1) as the set of linear polynomials $ax+b$ with $x^2 = -1$, i.e. with x interpreted as i. Indicate the generalisation: if $F[x]$ is the integral domain of polynomials $f(x)$ over the field F, then taking remainders on dividing $f(x)$ by a specified polynomial $g(x)$ gives a field — the field of polynomials {mod $g(x)$} — provided that $g(x)$ is irreducible in F.

CHAPTER 4

SETS

4.1. The basic concept of a set. The idea of a set is at the basis of all mathematics. As a description of the idea it is enough to say:

A set is a collection of well-defined objects thought of as a whole.

The qualification 'well-defined' is essential; the objects must be precisely specified. The specification may be by listing all the objects, or it may be the provision of a property or formula which describes some common characteristic of the objects. As a result we can adopt convenient *notations* for a set, writing a capital letter for the set itself and small letters for the objects of the set:

$$A = \{a, b, c, \ldots\} = \{a \mid a \text{ is } P\}.$$

The first is particularly appropriate when the objects are listed, the second when they are characterised by some formula or property P. In the second notation, the vertical line | is to be read 'such that'; A is the set of all objects a such that 'a is P', describing the common property.

The 'objects' which are the *members* or *elements* of the set may be entities of any kind. The basic concept of a set includes the idea that a set is composed of members and that members belong to the set. To denote that the object a is 'a member of' or 'belongs to' the set A, we write $a \in A$. Two examples illustrate. The ten digits used in the decimal system of writing rational numbers make up a set A; the members of A are digits, particular marks on the paper. A can be listed, or it can be described by a property:

$$A = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\} = \{a \mid a \text{ is a digit}\}.$$

The boys assembled in the Lower Third of a named school at a specified time compose a set B. The 'objects' are now boys; they are described by the property given and they can be listed (e.g. in the class register if it is well-kept). So Brown of the Lower Third belongs to B if he was present at the time, whereas Green, an absent Lower Thirder, does not. We can write Brown $\in B$ and Green $\notin B$.

Finally, the words 'thought of as a whole' indicate that the concept of a set is somewhat sophisticated. The objects may be easily comprehended but a collection of objects is a further and abstract idea; the totality is more than the sum of the parts. A boy is an 'object' easily recognised; the set of boys making up the Lower Third is a little more difficult. Or, to take another set of persons, even if we know what father, mother, aunts and uncles are, a set of them requires the idea of members of the same generation in a family.

Having said all this, we still have the task of postulating the properties we wish a set to possess. We might try to take the description of a set given above as a definition and then to deduce properties. This would not take us very far. It seems better, in the end, to take 'set' as a primitive (undefined) concept and 'member of' as a primitive (undefined) relation. We are then free to define other concepts, such as 'subset', in terms of them, and to lay down as axioms the properties we find we need. Set theory, though basic in mathematics, is relatively new, stemming from the work of Cantor (1845–1918). An axiomatic formulation is an even more recent development and the one followed here is essentially that of Zermelo (1871–1956). The set of Zermelo's axioms, suitably adjusted to the present approach, is shown formally in 15.3, but what the axioms attempt to do is quite easily described in general terms. A first axiom asserts that a set is completely fixed by specifying its members, so making possible the notation $A = \{a, b, c, \ldots\}$. Two axioms follow to permit sets to be built up from smaller ones and to be obtained by breaking down larger sets into subsets. Another pair of axioms is required to allow sums and products of sets. For two sets, summation yields their 'union' as the set of members belonging to one set *or* the other. One product is the 'intersection' of two sets, the set of members belonging to one set *and* the other. A second kind of multiplication gives the 'Cartesian product', the set got by selecting pairs of members, one from each set.*

Even these five axioms are not enough. Many sets are finite,

* The intersection is sometimes called the 'inner product' and the Cartesian product the 'outer product' of the two sets.

having n members, where n is a positive integer. Those quoted above are instances. But more usually we handle 'infinite' sets, e.g. the set of all integers or of all rationals. The five axioms are then found to be insufficient. Indeed, the step from the 'finite' to the 'infinite' is a tremendous one, and one of the utmost importance. The step is from the set of natural numbers $\{1, 2, 3, \ldots n\}$ to the set of all natural numbers $\{1, 2, 3, \ldots n, \ldots\}$. In adding ... to n, we are making a major conceptual advance into completely new and open country. It will occupy our attention later in this chapter.

Many sets, such as those of Chapter 2, have numbers as elements. A different kind of set is met in Chapter 3, a set of polynomials each member of which is itself a set (ordered sequence) of coefficients. Here is a case of a set of sets. A further instance is a set of 'matrices', where a 'matrix' is an ordered arrangement of entities (e.g. numbers) in rows and columns (Chapter 13). Still another kind is a set of operations and particularly of transformations (Chapters 6 and 7). There is no end to the different sorts of sets which can be specified. The examples below provide instances of sets of people, or other everyday entities.

A further notation is needed to indicate cases where one set is contained within another, a concept roughly corresponding to 'less than' for ordered numbers or magnitudes. Write $A \subseteq B$ for A as a *subset* of B if each element of A belongs also to B. This includes $A = B$, A identical with B, as the particular case where A and B contain precisely the same elements. Hence it is possible to write both $A \subseteq B$ and $B \subseteq A$, simply meaning $A = B$. Write $A \subset B$ for A as a *proper subset* of B where $A \subseteq B$ but $A \neq B$. Some examples illustrate:

(i) $J = \{\ldots -2, -1, 0, 1, 2, \ldots\} = \{n \mid n \text{ an integer}\}$ as the set of all integers. Then $J^+ = \{1, 2, 3, \ldots\} = \{n \mid n \text{ a positive integer}\}$ is a proper subset of J.

(ii) Six people sit down to dinner: the host and hostess and two married couples, X and X', Y and Y' respectively. The table is a rectangle and the host and hostess sit one at each end. The other four places are labelled (a), (b), (c), (d). The set S of all arrangements of the seating consists of 24 elements, listed on p. 91.

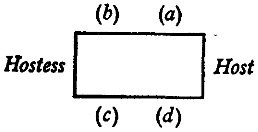
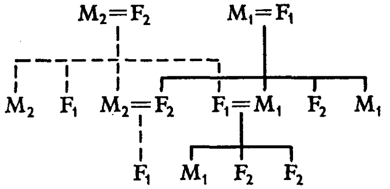


FIG. 4.1a

A proper subset S_1 of S consists of all arrangements in which each married couple is separated, i.e. X and X' do not sit together, neither do Y and Y'. There are 16 elements in S_1, as seen from the listing below. There are four arrangements (numbered 7, 9, 20 and 23) where the sexes alternate round the table: host, female, male, hostess, male, female. This is also a proper subset S_2 of S. The two subsets S_1 and S_2 overlap, having two elements in common (numbered 9 and 20). This subset S_3 of two elements comprises the alternative arrangements which the hostess probably has in mind: separating the sexes and the married couples. S_3 is a proper subset both of S_1 and of S_2.

List	*Seats*				*List*	*Seats*				*List*	*Seats*			
no.	(*a*)	(*b*)	(*c*)	(*d*)	*no.*	(*a*)	(*b*)	(*c*)	(*d*)	*no.*	(*a*)	(*b*)	(*c*)	(*d*)
1	X	X'	Y	Y'	9	X'	Y	X	Y'	17	Y	Y'	X	X'
2	X	X'	Y'	Y	10	X'	Y	Y'	X	18	Y	Y'	X'	X
3	X	Y	X'	Y'	11	X'	Y'	X	Y	19	Y'	X	X'	Y
4	X	Y	Y'	X'	12	X'	Y'	Y	X	20	Y'	X	Y	X'
5	X	Y'	X'	Y	13	Y	X	X'	Y'	21	Y'	X'	X	Y
6	X	Y'	Y	X'	14	Y	X	Y'	X'	22	Y'	X'	Y	X
7	X'	X	Y	Y'	15	Y	X'	X	Y'	23	Y'	Y	X	X'
8	X'	X	Y'	Y	16	Y	X'	Y'	X	24	Y'	Y	X'	X

(iii) In a tribe, all members belong to one class (upper class) or another (lower class). Marriage is permitted only between men and women of the same class; sons take the same class as their parents, daughters the other class. (The idea here may be the prevention of in-breeding.) The set S of 16 people, shown in a family tree, consists of three generations, the grandparents being an upper class couple and a lower class couple respectively. One proper subset S_1 consists of the three unmarried men; another S_2 of the five unmarried women. The marriages to be arranged pair off an element of S_1 with a suitable element of S_2 (M_1 with F_1, M_2 with F_2). The marriage laws are such that M_1 can marry his mother's sister or his mother's brother's daughter; he cannot



FAMILY TREE
M: Male F: Female =: Married
Subscripts 1: Upper Class 2: Lower Class

Fig. 4.1*b*

marry his father's sister (or his father's brother's daughter if he has one).

4.2. Operations on sets. Consider a given totality of elements and form all possible sets A, B, C, ... from the elements. There are two special (limiting) sets: the *universal set* U of all elements and the *empty set* ϕ of no elements. For completeness, these are included with other sets of which they form the *bounds*:

$$\phi \subseteq A \subseteq U \quad \text{for any set } A.$$

Three operations are defined. One is a *unary operation*, involving a single set A, and gives the complement A' of A as the set of all elements not in A. The other two are *binary operations*, involving a pair of sets A and B. One gives the union $A \cup B$ as the set of elements which are in A, in B or in both, the other the intersection $A \cap B$ as the set of elements which are in both A and B.

DEFINITION: *The* **complement** A' *of* A *is the set of those and only those elements which are* **not** *in* A. *The* **union** $A \cup B$ *of* A *and* B *is the set of those and only those elements which are in* A **or** *in* B *(or both). The* **intersection** $A \cap B$ *of* A *and* B *is the set of those and only those elements which are in* A **and** *in* B.

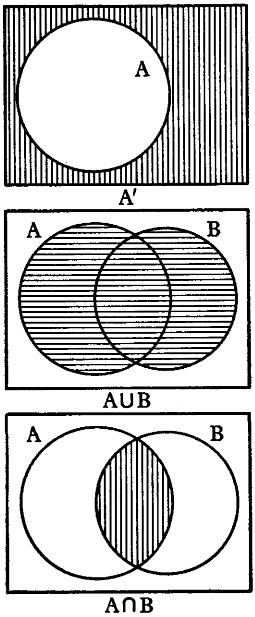


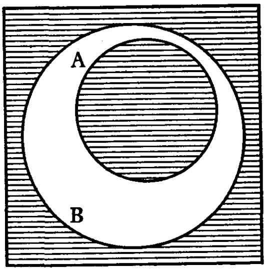
FIG. 4.2*a*

The symbols $\cup$ and $\cap$ may be read 'cup' and 'cap' respectively.

These operations on sets can be illustrated by drawing what are termed *Venn Diagrams.* The universal set U is represented by points within a rectangle, and sets A, B, ... are shown by points within circles drawn inside the rectangle. Fig. 4.2*a* illustrates the results of the operations of complement, union and intersection (shaded area in each case). Too much stress should not be laid on these diagrams; they are no more than helpful illustrations.

Set theory deals with these operations on sets and with certain relations which can be defined in terms of them. One useful relation is $A \cap B = \phi$, i.e. the intersection of the sets

A and B is empty. In this case, A and B have no elements in common; they are *disjoint sets*. The relation of *inclusion* $A \subseteq B$, already introduced, can now be expressed: $A \cap B' = \phi$, i.e. A and the complement of B are disjoint, as in Fig. 4.2*b* where A and B' are shaded. For example, if R is the set of all rational numbers, then the definition of a real number (2.4) divides R into two disjoint and exhaustive sets L and G: $L \cup G = R$ and $L \cap G = \phi$.



If $A \subseteq B$, then $A \cap B' = \phi$

FIG. 4.2*b*

It is evident that the operations (′, ∪, ∩) are closely linked with the verbal ideas of 'not', 'or', 'and' respectively. These are the words in bold in the definition above. This will be followed up in Chapter 5.

4.3. The operational rules for sets. Meanwhile, another line of thought is pursued. The two binary operations (∪, ∩) are similar in many respects to the ordinary arithmetic concepts of sums (+) and products (×). This is particularly so for sums and products of proper fractions: $\frac{1}{2}+\frac{1}{3}=\frac{5}{6}$ getting 'bigger' like $A \cup B$ and $\frac{1}{2}\times\frac{1}{3}=\frac{1}{6}$ getting 'smaller' like $A \cap B$. On this line of approach, the two bounds (ϕ, U) are similar to the numbers 0 and 1 respectively. The similarity is by no means complete and there are many divergencies. But it is striking enough to pursue. The suggestion is that ∪ can be re-written as +, ∩ as ×, ϕ as 0 and U as 1.

A question now presents itself in set theory: what rules are obeyed by the operations of complement, union and intersection? The rules are got directly from the definitions as set out formally, and in terms of the notation ′, ∪ and ∩, in 15.3. They are to be used in this form in most of set theory, as applied for example in Chapter 5. Here, we try out the rules with the proposed more familiar notation: (+, ×, 0, 1) instead of (∪, ∩, ϕ, U). The notation for complement is retained. The translation of the rules is given on p. 94.

The last three rules (8, 9 and 10) relate to complements and they are reasonable enough. The test of the notation of + for union and × for intersection is the comparison of the first seven rules with the corresponding operational rules of ordinary algebra (2.2). It is seen that the proposed notation passes the test fairly well but not completely. The seven rules above all look familiar, with the exception of rule 4 (both parts), rule 6(*a*) and rule 7(*b*). The exceptions are not minor ones; indeed, they are strange and might even appear silly.

The Operational Rules for Sets

Sets $A, B, C, \dots$ with 0 as the empty and 1 as the universal set

Rule	*Union* (+)	*Intersection* (×)
1. Closure	(*a*) $A+B$ is a set	(*b*) $A\times B$ is a set
2. Associative	(*a*) $A+(B+C)=(A+B)+C$	(*b*) $A\times(B\times C)=(A\times B)\times C$
3. Commutative	(*a*) $A+B=B+A$	(*b*) $A\times B=B\times A$
4. Idempotent	(*a*) $A+A=A$	(*b*) $A\times A=A$
5. Bounds	(*a*) $A+0=A$	(*b*) $A\times 1=A$
6.	(*a*) $A+1=1$	(*b*) $A\times 0=0$
7. Distributive	(*a*) $A\times(B+C)=A\times B+A\times C$	(*b*) $A+(B\times C)=(A+B)\times(A+C)$
8. Complements	(*a*) $A+A'=1$	(*b*) $A\times A'=0$
9.	(*a*) $(A+B)'=A'\times B'$	(*b*) $(A\times B)'=A'+B'$
10. Involution	$(A')'=A$	

The algebra of sets, therefore, is not the same as the algebra of ordinary numbers. It is an example of what is known as *Boolean Algebra*, after Boole (1815–64). We have a choice here on the notation to adopt and it is a choice which appears elsewhere (e.g. in matrix algebra). We can (as in 15.3) make use of new and strange symbols, particularly $\cup$ for union and $\cap$ for intersection of sets. In this way, we separate Boolean Algebra completely from ordinary algebra and we choose to ignore the similarities between the two. On the other hand (as here), we can retain the symbols we are used to, writing + for union and × for intersection of sets. In this case, we depend on the similarities between Boolean and ordinary algebra. At the same time, we have to remember that not all the familiar rules of + and × are obeyed when these symbols are applied to sets.

It is not easy to make the choice. Generally speaking, however, we find much advantage in sticking to the familiar notation. We economise in symbols and, for most of the time, we are in well-known territory. But we have to clear our minds of the idea that all the rules of ordinary algebra have to be obeyed. It is useful to consolidate this position here and now. Later (in matrix algebra) we will find that the same problem arises; the established notation does make use of + and ×, despite the fact that the operational rules are not all valid.

4.4. Boolean Algebra. If the algebraic operations with certain entities satisfy the set of rules set out in 4.3, for appropriate definitions of complement, union and intersection, then the algebra is Boolean. It is the algebra obeyed by sets, as developed above. It is, moreover, an algebra with a considerable range of application, as in Chapter 5 where the complement, union and intersection of sets are related to verbal statements in 'not', 'or', and 'and'. This is, however, not the only Boolean Algebra which can be devised, as a simple example demonstrates. As in 2.8, consider a set of two elements $\{0, 1\}$ subject to the operations:

Complements: $0'=1$ and $1'=0$

Addition:

+	0	1
0	0	1
1	1	1

Multiplication:

×	0	1
0	0	0
1	0	1

All the operational rules of 4.3 are then satisfied (see 4.9 Ex. 11).

A more detailed examination of the operational rules of Boolean Algebra is now made, to see to what extent they are in line with those of ordinary algebra, and in what respects they differ. We accept as familiar and/or very reasonable all the rules of 4.3, with three exceptions. Those which are 'different' in one way or another are rules 4 and 6(a) on the one hand, and the distributive rule 7 on the other.

When extended (as it obviously can be) to several terms, rule 4 gives:

$$A+A+A+\ldots(n \text{ terms})=A \quad \text{and} \quad A\times A\times A\times\ldots(n \text{ terms})=A.$$

On the ordinary rules of algebra, we expect nA and A^n respectively. The difference here is that the Boolean rule is the simpler. Similarly, rule 6(a) is strange: $A+1=1$, i.e. 'adding' anything to the universal set 1 leaves it unchanged. This is, however, exactly parallel to the rule 6(b): $A\times 0=0$, i.e. 'multiplying' the empty set by anything leaves it unchanged. We are used to the second, a property of zero; but we are not used to the parallel property applied to unity. The difference here is that the Boolean rule is the more symmetrical.

The feature of symmetry in Boolean Algebra is even more remarkable in rule 7, the distributive rule. One part is familiar and acceptable, the distribution of × over + : $A\times(B+C)=A\times B+A\times C$. The other part is strange, the distribution of + over × :

$$A+(B\times C)=(A+B)\times(A+C).$$

For any usual system of numbers, this second part of the rule is just not true, e.g. $2+(3\times 4)\neq(2+3)\times(2+4)$. For sets, and Boolean Algebra generally, both parts are true and the one is exactly parallel to the other, obtained by interchanging the operations ($+$ and $\times$). Hence the leading question: which should we naturally expect, one distributive rule or two?

Pausing to take stock, we observe that the Boolean rules are indeed completely symmetrical. They have the property of *duality*: if the operations $+$ and $\times$ are interchanged, and if (at the same time) 0 and 1 are interchanged, then any one of the rules is transformed into its pair. So, for example, $A+0=A$ changes into $A\times 1=A$, and $A+1=1$ into $A\times 0=0$. It is in this way that the one distributive rule transforms into the other. Surely, this symmetry or duality is a very desirable feature of a system. Boolean Algebra has it; ordinary algebra does not. It can only be concluded that, in ordinary algebra, we do not realise what we are missing. Without knowing it, we have put up with an unsymmetrical system; and we have more complicated rules of repeated addition and multiplication than we need:

$$A+A+A+\ldots=nA;\ A\times A\times A\times\ldots=A^n.$$

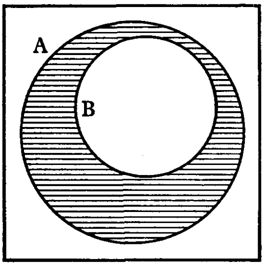
Boolean Algebra is neater: more symmetrical and simpler

$$A+A+A+\ldots=A;\ A\times A\times A\times\ldots=A.$$

However, there is something to be put in the scales on the other side. Boolean Algebra, like ordinary algebra, has identities 0 and 1. The rules of Boolean Algebra, unlike those of ordinary algebra, say nothing about inverses (negatives and reciprocals), or even about cancellation. This need cause no concern as far as negatives and hence differences go. Provided only that $B\subseteq A$, we can write the difference $A-B$ between the sets A and B. First, denote $B'=1-B$, i.e. B' is also the difference between the universal set and B. Then, note:

$$AB'=A(1-B)=A-AB=A-B \quad (B\subseteq A)$$

all represent the same thing, the set consisting of those and only those elements of A which are not in B. This is illustrated in the Venn Diagram of Fig. 4.4.



A–B
FIG. 4.4

Hence, the lack in Boolean Algebra is effectively reduced to the following: there are no reciprocals (and hence no division) and cancellation is not valid. Given a set A, there is *no* set B such that $A \times B = 1$.* Moreover, it is *not* true that: if $A \times B = 0$, then either $A = 0$ or $B = 0$. It is true that, if either A or B is 0 (empty), then $A \times B$ must also be 0 (empty). It is the converse which fails. Indeed:

$$A \times B = 0 \quad \text{implies} \quad A \text{ and } B \text{ disjoint.}$$

In Boolean Algebra, there are no reciprocals; even worse, there are divisors of zero, i.e. any disjoint sets have a product (intersection) which is zero (empty).

While this is a definite lack in Boolean Algebra, it is certainly not the only system with the defect. There are even systems of numbers equally defective, e.g. the rather sophisticated algebra of integers (mod n) where n is not prime (see 2.7). So, in the algebra of $\{0, 1, 2, 3\}$ (mod 4), we have $2 \times 2 = 0$. It all depends whether we are happy without reciprocals and a cancellation rule; if so, Boolean Algebra has much to recommend it.

4.5. Counting sets. Having established the rules for operating with sets, we can turn to another promising question: how do we count how many elements a set has? This is indeed a basic, if not primitive, idea. We find, however, that we are soon on unfamiliar ground. The concept of counting leads on to that of the infinite, and here we need to be both cautious and precise.

It can be agreed at the outset that counting has its *ordinal aspect*, i.e. ticking off in sequence 1, 2, 3, ..., and that this aspect has a very close link with the ordered set J^+ of positive integers. This is, however, not the basic property of counting. To count a set in this ordinal way, we need to be able to arrange the set (or at least part of it) in sequence; we do not know whether this can be done at all (see the 'axiom of choice', 4.8 below) or, if it can be done, whether it can be done uniquely. A much more fundamental view of counting is from its *cardinal aspect*. A set is simply a collection of elements, no idea of order being involved. We want to count it.

The essential idea, from which counting derives, is that of *one-one correspondence*, a most far-reaching concept in mathematics. In

* Except (as always) that 1 is the reciprocal of itself: $1 \times 1 = 1$.

everyday terms, one-one correspondence is simply 'matching'. When we say that two sets have the same number of elements, the same count, we mean that the elements of one set can be matched off against those of the other. For example, a set of three eggs in a basket can be matched against a set of three sticks on the ground: each egg paired off with one stick and conversely. The pairing may also be done indirectly, e.g. if the basket of eggs is in one place and the sticks in another. The eggs can be matched off against a set of three fingers and then the sticks matched off against the same set of fingers. All these sets have the same number of elements; the fact, that we say that the number is 3, is incidental at this stage. So:

DEFINITION: *Two sets A and B can be put into* **one-one correspondence** *written $A \sim B$, if each element of A can be associated with just one element of B and conversely. A and B are then* **equi-numerous** *and have the same* **cardinal number.**

There is nothing in this definition to imply that the sets A and B are finite. The illustration of three eggs and three sticks is an instance of a finite (cardinal) number. The concept of one-one correspondence, and of cardinal numbers, applies to infinite sets just as well, as illustrated by the following example.

The set of even positive integers may be thought to be half as numerous as the set of all positive integers. For a finite set of each, this may be so, roughly. For example, $\{2, 4, 6, 8\}$, the even integers less than 10, has 4 elements; $\{1, 2, 3, \ldots 9\}$, the integers less than 10, has 9 elements. Any matching of these sets leaves some left over in the second and 'larger' set. All such 'end effects', left-overs in the matching process, disappear when *all* even integers and *all* integers are matched. There is a simple one-one correspondence: to each integer n match the even integer $2n$ and conversely. The sets of even integers and of all integers are equi-numerous and have the same (infinite) cardinal number.

On this definition, each and every set has its cardinal number. If there are no sets to be matched with it, the cardinal number is unique. But otherwise (and generally) all sets which can be put in one-one correspondence ($A \sim B$) have the same cardinal number. Note that counting and cardinal numbers are confined (at least at this stage) to sets. We must not slip into the habit of saying that (for example) 3 lbs. of sugar or 3 feet of rope are equi-numerous with 3 eggs or 3 sticks. The sugar and the rope are not sets and the number 3 attached represents a more developed idea (measurement) than that of counting.

The equi-numerous relation $A \sim B$ is a first case of an 'equivalence relation', met later in a more general context (Chapter 7). As such, the following properties of equi-numerous sets A, B, C, ... are of interest:

(1) *Reflexive:* $A \sim A$, (2) *Symmetric:* If $A \sim B$, then $B \sim A$,
(3) *Transitive:* If $A \sim B$ and $B \sim C$, then $A \sim C$.

The arithmetic of cardinal numbers, a, b, ..., can then be constructed in terms of the equi-numerous sets to which each corresponds. Only two simple pieces of the arithmetic need be exhibited here.

Let a be the cardinal number of the set A (and all equi-numerous sets) and let b be the cardinal number of the set B (and all equi-numerous sets). *Define* the relation $a \leqslant b$ (a less than or equal to b) as holding if: $A \sim$ subset of B. But: $a = b$ if and only if $A \sim B$. It does *not* follow, however, that if $A \sim$ proper subset of B, then $a < b$. It is still possible that this proper subset of B is equi-numerous with B itself and, hence, that $A \sim B$ and $a = b$. Ample illustration can be given of cases where a set can be put into one-one correspondence with a proper part of itself. It is enough here to note the example given above; the set of even positive integers is a proper subset of the set of all positive integers, and the two sets are equi-numerous. Hence, as long as we can put a set A into one-one correspondence with a subset of B (proper or not), we can write $a \leqslant b$ for the corresponding cardinal numbers. It is only possible to write $a < b$, if $a \leqslant b$ and if we can establish otherwise that $a = b$ is not true.

Next, *define* the sum $(a+b)$ of two cardinal numbers as the cardinal number of the set $A+B$ formed as the union of two *disjoint* sets A and B with cardinal numbers a and b respectively. This is straightforward; the need for A and B to be disjoint if the numbers of elements are to be added is clear enough. What does need proof, however, is that the definition of $(a+b)$ is independent of the choice of A and B from all the possible disjoint sets with the cardinal numbers a and b. The proof, fortunately, is easy. Suppose A_1 and B_1 are another pair of disjoint sets with cardinal numbers a and b respectively. Then $A \sim A_1$ and $B \sim B_1$. Combining these two one-one correspondences, we deduce that $(A+B) \sim (A_1+B_1)$. Hence the cardinal number $(a+b)$ is uniquely defined, from a and b.

It is possible to develop all the properties and operational rules of cardinal numbers from the definition in this way, i.e. without any reference whatever to the ordered set of positive integers. However, sooner or later, we wish to associate cardinal numbers and integral numbers, the one defined here in relation to equi-numerous sets, the other developed on the lines of 2.6. The sets, of three eggs and of three sticks, have the same cardinal number 'three'; we cannot hide the fact that this is equivalent in some way to the positive integer 3. Consequently, instead of proceeding further with cardinal numbers in the abstract, we cut a corner and use the ordered sequence of positive integers J^+ as the yardstick for counting. We lose some fine distinctions but we get to usable results more quickly. We must remember, however, that the essence of counting is the one-one correspondence; ordering is subsidiary.

4.6. Finite sets. The order of the set J^+ of positive integers is such that $m < n$ (or $n > m$) means that m is before n in the order and that the difference $(n-m)$ is defined as a positive integer of J^+. So 'less than' or 'greater than' is simply a property of order, of a positive difference. Denote a segment or *section* S_n of J^+ as the subset

$$\{1, 2, 3, \ldots n\} = \{p \mid p \leqslant n\},$$

i.e. all positive integers before n (including n itself) in the order of J^+. Then a set A which is equi-numerous with S_n for *some* integral n can be called finite:

DEFINITION: *A set A is* **finite** *if* $A \sim S_n$ *for some section* S_n *of* J^+. We can then write $A = \{a_1, a_2, a_3, \ldots a_n\}$. This follows from the fact that the elements of A can be associated (in one-one correspondence) with the integers 1, 2, 3, ... n, i.e. they can have these integers attached as subscripts.

It is tempting to say, right away, that A has n elements, i.e. that A's count or cardinal number is n, as given by the section S_n with which A is linked. The temptation must be resisted. It has not yet been shown that the linking of A with n is unique; only that it exists for some n. We need the important result:

THEOREM: *If a* **finite** *set A is a* **proper subset** *of another set B, then $A \sim B$ is impossible; A and B cannot be equi-numerous.*

To prove, suppose that $A \sim B$ and that $A = \{a_1, a_2, \dots a_n\}$. Hence, B consists of the same elements and (at least) one other a_{n+1} (since $A \subset B$). We have to show that this is impossible. The proof is by mathematical induction (see 2.6), i.e. we show that it is impossible for $n = 1$ and then, if taken as impossible for $(n-1)$, it is also impossible for n.* The first part is easy: If $n = 1$, A is a_1 alone and B has (at least) a_1 and a_2 as elements. No one-one correspondence for $A \sim B$ is possible. For the second part, take $A \sim B$ as impossible for $(n-1)$. For n, assume $A \sim B$ with $A = \{a_1 a_2 \dots a_n\}$ and B as the same elements *plus* a_{n+1} at least. Since $A \sim B$, there is a one-one correspondence between A and B, which can always be arranged (if necessary by switching two elements in A) so that a_n in A corresponds to a_{n+1} in B. Remove a_n from A, a_{n+1} from B. There is still a one-one correspondence between $\{a_1 a_2 \dots a_{n-1}\}$ from A and the same elements *plus* a_n in B. It is this which we have taken as impossible. So, if impossible for $(n-1)$, it is impossible for n. The proof by induction is complete. Q.E.D.

One consequence of the theorem is immediate: a finite set A cannot be equi-numerous with two sections S_n and S_m of J^+ $(n \neq m)$. Suppose $m < n$, so that S_m is finite and a proper subset of S_n. If $A \sim S_n$ and $A \sim S_m$ is possible, then $S_m \sim S_n$, which is ruled out by the theorem. Hence

$$A \sim S_n \sim S_m \quad \text{if and only if } n = m.$$

It follows that the integral n of S_n, in the definition of A finite, is unique. In writing $A = \{a_1, a_2, \dots a_n\}$ for a finite set, the integer n is unique; it is the cardinal number of A. All equi-numerous sets can be written:

$$A = \{a_1, a_2, \dots a_n\};\ B = \{b_1, b_2, \dots b_n\};\ \dots$$

the *cardinal number n* being unique, the same for each.

Since we can now count a finite set, we can proceed to operate with the various counts of finite sets A, B, C, ... Addition of numbers (of elements) is the most important of the operations. Write $n(A)$ and $n(B)$ as the number of elements in the sets A and B

*** This form of mathematical induction, when spelled out, implies: impossible for $n = 1$, hence impossible for $n = 2$, hence impossible for $n = 3$, ... and so impossible, generally, for any n.**

respectively. What is $n(A+B)$ where $A+B$ is the union of A and B? By the definition of addition of cardinal numbers (4.5):

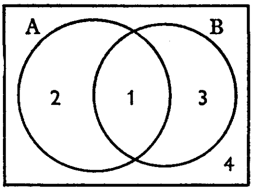
$$n(A+B) = n(A) + n(B) \quad \text{if } A \text{ and } B \text{ are disjoint.}$$

We can add the numbers of elements in two sets, if they are disjoint, i.e. if they do not overlap. So far, so good (and so obvious). The extension to the result when A and B do overlap is a matter of picking out suitable disjoint sets before adding numbers of elements. The Venn Diagram of Fig. 4.6*a* shows disjoint sets making up A, B and $(A+B)$. Here AB is written for $A \times B$ or $A \cap B$. So:

$$n(A+B) = n(AB) + n(AB') + n(A'B)$$
$$n(A) = n(AB) + n(AB')$$
$$n(B) = n(AB) + n(A'B).$$

Hence:

$$n(A+B) = n(A) + n(B) - n(AB) \quad \text{......(1)}$$



1=AB 2=AB′ 3=A′B 4=A′B′

FIG. 4.6*a*

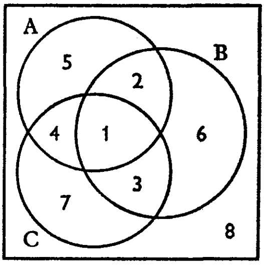
Again, this is sensible enough for, in writing $n(A)+n(B)$, we count the common elements, $n(AB)$ in number, twice over.

Two sets A and B divide the totality of elements into four disjoint and exhaustive subsets, numbered 1, 2, 3 and 4 in the Venn Diagram. The same process for three sets A, B and C requires eight disjoint and exhaustive subsets of the totality of elements, as shown by the Venn Diagram of Fig. 4.6*b*. Exactly as before, it follows that:

$$n(A+B+C)$$
$$= n(A) + n(B) + n(C) - n(AB) - n(BC) - n(AC) + n(ABC) \quad \text{......(2)}$$

The result (2) can be regarded as repeated applications of (1). More directly, if the numbers on the right-hand side of (2) are given, then the numbers in all other sets and combinations of sets can be obtained by a process of differencing. In particular, the numbers in the sets labelled 1, 2, 3, ... 8 in the Venn Diagram are obtained in sequence:

$$n(ABC) = \text{given}$$
$$n(ABC') = n(AB) - n(ABC) \quad \text{given}$$
$$n(BCA') = n(BC) - n(ABC) \quad \text{given}$$
$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$



1=ABC 2=ABC′ 3=BCA′
4=ACB′ 5=AB′C′ 6=BA′C′
7=CA′B′ 8=A′B′C′

Fig. 4.6*b*

It is required that each of these differences is non-negative, i.e. that the given numbers are consistent. The following examples illustrate two cases where the process turns out to be consistent and inconsistent respectively:

(i) A market research team interviews 100 people, asking each whether he smokes any or all of the items, A: cigarettes, B: cigars, C: pipe tobacco. For one individual, the answer could be none, any one of them, any pair of them or all three of them. The team returns the following numbers of over-lapping categories:

Category		*No.*	*Category*		*No.*
ABC	All three	3	A	Cigarettes	42
AB	Cigarettes and cigars	7	B	Cigars	17
BC	Cigars and pipe	8	C	Pipe	27
AC	Cigarettes and pipe	12		Total cases	100

It is required to unscramble these returns, by elimination of overlapping, and to find (e.g.) the number of people who smoke cigarettes and/or cigars but not a pipe. In particular, the number of nonsmokers is required. In the course of doing this, it can be determined that the returns are consistent. The Venn Diagram of Fig. 4.6c has eight disjoint categories and the number in each is in sequence:

$$n(ABC) = 3 \quad \text{(given)}$$
$$n(ABC') = n(AB) - n(ABC) = 7 - 3 = 4$$
$$\cdots\cdots\cdots\cdots\cdots$$

Consequently, entering the numbers in the diagram, we find the number of cigarette and cigar smokers who do not smoke a pipe: $26 + 4 + 5 = 35$. The number of non-smokers: $n(A'B'C') = 38$. This can be checked by means of the formula (2):

$$n(A + B + C) = 42 + 17 + 27 - 7 - 8 - 12 + 3$$
$$= 62$$

and $\quad n(A'B'C') = 100 - n(A + B + C) = 38.$

(ii) Suppose the team made the returns as in (i), except for the entries:

BC Cigars and pipe 13
AC Cigarettes and pipe 18.

There is still no reason, on the face of it,

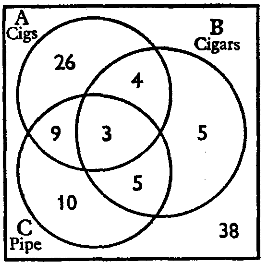


Fig. 4.6c

to expect inconsistent returns. The split into eight disjoint categories proceeds as before:

$$n(ABC)=3 \qquad n(ABC')=7-3=4 \qquad n(BCA')=13-3=10$$
$$n(ACB')=18-3=15 \qquad n(AB'C')=42-(3+4+15)=20$$
$$n(BA'C')=17-(3+4+10)=0$$
$$n(CA'B')=27-(3+10+15)<0 \text{ inconsistent}$$

i.e. the returns on pipe smoking are inconsistent. The figures are entered on a Venn Diagram as in (i), until an inconsistency (negative residual) appears.

4.7. Countably infinite sets. An *infinite set* is simply a set which is not finite, i.e. a set which cannot be put into one-one correspondence with any section of the set J^+ of positive integers. As a corollary of the theorem of 4.6, it follows that the *set J^+ is infinite*. For, if J^+ is finite, then $S_n \sim J^+$ for some section S_n of J^+. But S_n is a proper subset of J^+; this contradicts the theorem.*

The set $J^+=\{1, 2, 3, \dots n, \dots\}$, ordered in sequence, is the simplest kind of infinite set and (intuitively) it must have the 'lowest' count of all infinite sets. However this may be, consider infinite sets $A, B, C, \dots$ each of which can be put into one-one correspondence with J^+ $(A \sim B \sim C \sim \dots \sim J^+)$. They are called denumerable or *countably infinite*. They all have the same cardinal number which can be written d (for denumerable). From the definition of 4.5, it follows that $n \leqslant d$ for any finite cardinal number (positive integer) n since a subset of J^+ (i.e. S_n)$\sim S_n$. Moreover, $d \neq n$, since $d=n$ implies that there is a one-one correspondence between J^+ and S_n or that J^+ is finite. Hence $n<d$ for any integral n. Finally, if A is countably infinite, the one-one correspondence with J^+ means that the elements of A can be denoted with subscripts 1, 2, 3, ... n, ... All this can be summed up:

DEFINITION: *A set A is* **countably infinite** *if $A \sim J^+$. It can be denoted $A=\{a_1, a_2, a_3, \dots a_n, \dots\}$ and its cardinal number is $d>n$.*

Finite and countably infinite sets together can be described as *countable*. To tidy up the notation and summarise the position, we can say (as can easily be established) that a set A is countable if and only if it can be written $A=\{a_1, a_2, a_3 \dots\}$ with distinct elements

* Strictly, an axiom is required in set theory to ensure that infinite sets do exist. The axiom (see 15.3) can be simply that all natural numbers form a set J^+.

having distinct subscripts. A is finite if there is a last element a_n, the number of elements being n. A is countably infinite if the sequence continues indefinitely, the number of elements being d.

The theorem of 4.6 holds only for finite A. If A is infinite and a proper subset of B, then $A \sim B$ is still a possibility. It *may be* that an infinite set can be put into one-one correspondence with a proper subset of itself. This is 'may be'; it has not been established that it 'must be'. However, for countably infinite sets, it is 'must be', in virtue of the result:

THEOREM: *Any countably infinite set A can be put into one-one correspondence with some proper part A_1 of itself.*

This apparent paradox (sometimes known as the Paradox of Galileo, after Galileo, 1564–1642) is easily illustrated. The set $J^+ = \{1, 2, 3, \ldots\}$ is countably infinite, as is the set $\{2, 4, 6, \ldots\}$ of even positive integers which can be put into one-one correspondence with J^+. Hence J^+ can be put into one-one correspondence with a proper subset of itself. The general proof of the theorem is also simple. Let

$$A = \{a_1, a_2, a_3 \ldots\}$$

and write a proper subset $A_1 = \{a_2, a_3, a_4, \ldots\}$, i.e. A itself without the first element a_1. There is an obvious one-one correspondence (a_1 with a_1, a_2 with a_3, ...) between A and A_1. This proves the theorem. All that has been demonstrated (or indeed illustrated) is that there is *some* countably infinite set contained *within* any given countably infinite set. It is clear, however, that there are many such, e.g. within J^+ the following proper subsets are all countably infinite:

$$\{2, 3, 4, \ldots\}, \{2, 4, 6, \ldots\}, \{1, 3, 5, \ldots\}, \{2, 4, 8, \ldots\}.$$

The range of countably infinite sets is remarkable. The following very broad result demonstrates how they can be 'manufactured' one from another:

THEOREM: *The union of a countable set of countable sets is itself countable.*

Proof: let $A_1 = \{a_{11}, a_{12}, a_{13}, \ldots\}$, $A_2 = \{a_{21}, a_{22}, a_{23}, \ldots\}$, $A_3 = \{a_{31}, a_{32}, a_{33}, \ldots\}$, ... be the given sets. Throw all the elements together to get the union $A = A_1 + A_2 + A_3 + \ldots$ and A contains a_{rs}, the sth element of A_r, for any integral r and s. There is one element a_{rs} with $r+s=2$, two elements a_{rs} with $r+s=3$, ... and a finite number of elements with $r+s=n$ (any positive integer). Put the elements a_{rs} into this sequence, re-labelling them with one subscript only. The elements are thus countable. It does not matter whether A_1, A_2, A_3, ... are disjoint or not; if not, the suppression of the repeated elements does not affect essentially the sequence of the elements of A. Q.E.D.

As special cases of the theorem, we can write the union of a finite number of finite sets (itself finite), or the union of a finite number of countably infinite sets (itself countably infinite), or the union of a countably infinite number of countably infinite sets (still countably infinite). This helps to explain why countably infinite sets are of such frequent occurrence. J^+ (positive integers) is countably infinite, and so is J (all integers) which is the union of two countably infinite sets (the positive, the negative integers) together with zero. Moreover, both the set of rationals R, and the set $F[x]$ of all polynomials $f(x)$ with rational coefficients and undefined x, are countably infinite. Consider these in turn, in order to see how (following the lines of the general proof above) the whole set can be arranged in sequence.

Within R, the positive rational numbers can be put into sequence:

$$\frac{1}{1},\ \frac{2}{1},\ \frac{1}{2},\ \frac{3}{1},\ \frac{2}{2},\ \frac{1}{3},\ \frac{4}{1},\ \frac{3}{2},\ \frac{2}{3},\ \frac{1}{4},\ \ldots$$

Every positive rational fits into its place. The rational p/q is the nth element in the sequence, where $n=q+\frac{1}{2}(p+q-1)(p+q-2)$. The elimination of duplication (e.g. 2/2 or 4/2) makes no difference; a countably infinite sequence remains. The inclusion of the negative rationals (and zero) does not affect the result; two countably infinite sets are combined. R is arranged as a countably infinite sequence.

The set $F[x]$ of polynomials with rational coefficients is an example of a countably infinite set of countably infinite sets. Consider, first, polynomials with integral coefficients. Those of zero degree can be put in sequence:

$$0,\quad +1,\quad -1,\quad +2,\quad -2,\quad +3,\quad -3,\ \ldots$$

Adding the linear polynomials (of degree one), we get a square array:

$$\begin{array}{cccccc}
0 & +1 & -1 & +2 & -2 & \ldots \\
x & x+1 & x-1 & x+2 & x-2 & \ldots \\
2x & 2x+1 & 2x-1 & 2x+2 & 2x-2 & \ldots \\
3x & 3x+1 & 3x-1 & 3x+2 & 3x-2 & \ldots \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{array}$$

A single sequence is obtained by running up and down diagonals:

$$0,\ x,\ +1,\ -1,\ x+1,\ 2x,\ 3x,\ 2x+1,\ x-1,\ +2,\ -2,\ x+2,\ \ldots$$

This is made into an array by adding the x^2 terms for quadratic polynomials (of degree 2), again turned into a sequence, and so on. In the end the whole set of polynomials appears as a single sequence; it is countably infinite. Finally, a polynomial with rational coefficients is a rational multiple of a polynomial with integral coefficients. The countably infinite set of the latter is repeated a countably infinite number of times (once for each rational) to give $F[x]$. Hence, $F[x]$ is countably infinite.

4.8. Transfinite arithmetic. The arithmetic of finite cardinal numbers is simply the arithmetic of the positive integers. A new number is to be added, the cardinal number d corresponding to countably infinite sets and shown after the sequence of finite integers:

$$1,\ 2,\ 3,\ \ldots\ n,\ \ldots\ d \quad \text{(where all } n<d\text{)}.$$

The arithmetic can be appropriately extended, as will be done here for the operation of addition. Again something new is to be expected. From 4.5, the formal process of adding two cardinal numbers (infinite as well as finite) is no problem. It is only a matter of forming the union of appropriate disjoint sets and of writing the cardinal number of the union. The difficulty, rather, lies in the interpretation of this transfinite arithmetic. One point is evident; since the arithmetic of cardinal numbers reflects operations with sets, it is to be expected that it will display features of Boolean rather than ordinary algebra.

As a consequence of the second theorem, of 4.7, it follows that, for any positive integer n and the cardinal number d:

$$d+n=d;\ d+d=d;\ d+d+d+\ldots=d. \qquad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

There can be a countably infinite set of d's in the left-hand side of the last result of (1). To prove, it is only necessary to write appropriate *disjoint* sets, to form their union and to use the theorem.

This is not the end of the story. The 'infinite' is not a simple concept; it is found to have a structure. Despite the fact that so many infinite sets are countable, it is easy enough to find one that is not. The set R^* of real numbers contains within it many proper subsets which are countably infinite, e.g. the set R of rationals and the set

$R(\sqrt{2})$ obtained by adjunction of $\sqrt{2}$ as in 2.3 (4.9 Ex. 18). But the real numbers are so thick on the ground that, no matter how many countably infinite subsets are removed, there are always as many and more left. The result, due to Cantor (1845–1918), is:

THEOREM: *The set R^* of real numbers is not countably infinite.*

Proof: start with all real numbers x between 0 and 1 $(0<x<1)$. Each of them can be written as a decimal, in general not terminating. Suppose that the real numbers are countably infinite so that they can be written in a sequence: $x_1, x_2, x_3, \ldots$. In decimal form:

$$\begin{aligned} x_1 &= 0\,.\,a_{11}a_{12}a_{13}\ldots \\ x_2 &= 0\,.\,a_{21}a_{22}a_{23}\ldots \\ x_3 &= 0\,.\,a_{31}a_{32}a_{33}\ldots \\ &\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot \end{aligned}$$

where all the a's are digits from the set $\{0, 1, 2, \ldots 9\}$. Consider the diagonal of digits: $a_{11}, a_{22}, a_{33}, \ldots a_{nn}, \ldots$ which is countably infinite. Form a new sequence of digits: $b_1, b_2, b_3, \ldots b_n, \ldots$ where

$$b_n=1 \quad\text{if}\quad a_{nn}=0 \quad\text{and}\quad b_n=a_{nn}-1 \quad\text{if}\quad a_{nn}\neq 0.$$

Then each of the b_n is different from the corresponding a_{nn}. Write

$$b=0\,.\,b_1b_2b_3\ldots$$

which is a real number (between 0 and 1). However, b is different from each of the real numbers $x_1, x_2, x_3, \ldots$, i.e. it differs from x_n at least in the nth decimal place (since $b_n \neq a_{nn}$). This contradicts the assumption that the real numbers (between 0 and 1) are countably infinite, all being comprised in the sequence $x_1, x_2, x_3, \ldots$. Hence the real numbers between 0 and 1 are not countable. Similarly, the (double) set of real numbers between -1 and 1 is not countable.

To extend to the whole set of R^*, it is only necessary to get a one-one correspondence between the set of real numbers $-1<x<1$ and the set of all real numbers y. Such a correspondence is provided by $y=x/(1-x^2)$ $(-1<x<1)$. See 4.9 Ex. 19 and 20. It follows that the whole set R^* and the real numbers $(-1<x<1)$ have the same cardinal number. Write it c, so that $d \leqslant c$ since there is a proper subset of R^* (e.g. the positive integers) which is in one-one correspondence with J^+. But $c \neq d$ since R^* is not countable. Hence, as a definition and summary of results obtained:

DEFINITION: *The non-countable set R^* of real numbers has the* **cardinal number** $c > d$.

In the course of the proof above, it was established that the infinite set R^* could be put into one-one correspondence with a proper subset of itself, the real numbers $(-1 < x < 1)$. The set R^* with cardinal number c contains within itself a part (a proper subset) also with the cardinal number c.

Transfinite arithmetic is now extended further to include in sequence:

$$1, 2, 3, \dots n, \dots d, c \quad (\text{all } n < d < c).$$

At this stage, having shown the possibility of more than one infinite number, we can leave the development to those interested in this fascinating but (except for the main ideas) not very practical subject. As far as addition is concerned, it can be shown that the results (1) extend:

$$c + n = c;\ c + d = c;\ c + c + c + \dots = c \quad \dots\dots\dots\dots\dots\dots(2)$$

It would appear, from (1) and (2), that the sum of two or more infinite numbers is simply the greatest of the numbers. This is, in fact, the case; and for products as well as sums. It is a reflection of one of the rules of Boolean Algebra: $A + U = U$ where U is the universal set. Replace U by an infinite cardinal number and A by a lower cardinal number, and the main result of transfinite arithmetic for sums is obtained.

There is still some tidying up to do. A finite set is defined (4.6) as one which can be put into one-one correspondence with $\{1, 2, 3, \dots n\}$ for some natural number n. This *inductive* definition of the finite makes use of the characteristic feature (mathematical induction) of the natural numbers. A finite set is such that its members can be paired off with 1, 2, 3, ... until some n is reached and there is nothing left over. The corresponding inductive definition of an infinite set is by negation: an infinite set is not finite, not 'countable' against any natural number n.

The theorem of 4.6 states that a finite set cannot be put into one-one correspondence with a proper subset of itself. There is a property here which can be given a convenient label:

DEFINITION: *A set is* **reflexive** *if it can be put into one-one correspondence with a proper subset of itself.*

The theorem on finite sets can then be re-stated:

THEOREM: *A finite set is not reflexive.*

Now put up for examination the converse:

CONVERSE THEOREM: *A non-reflexive set is finite.*

We have not yet proved this; clearly we would very much like to do so. For the moment, let us take the result as established and see what further light is thrown on the nature of the infinite.

On the inductive definition, an infinite set is one which is not finite; it may still be reflexive or non-reflexive. On the converse theorem, a set which is not reflexive must be finite, cannot be infinite. So the possibility of an infinite and not-reflexive set is ruled out; an infinite set must be reflexive. By 4.7, a countably infinite set is reflexive; we now see that all infinite sets are reflexive. The characteristic property of the infinite is reflexivity. With infinite sets it becomes possible that two sets of different 'sizes' can be made to correspond member by member, as the set of even integers corresponds to the 'larger' set of all integers. We have the tidy result:

(i) A *finite set* is both inductive (countable against n) and non-reflexive (not in one-one correspondence with a proper subset).

(ii) An *infinite set* is both non-inductive (not countable against n) and reflexive (in one-one correspondence with a proper subset).

Indeed, an alternative and equivalent definition of the finite and infinite can be given: an infinite set is one which is reflexive and (by negation) a finite set is one which is not infinite.

The problem remains: can we prove the converse theorem. At first sight it seems rather an easy matter. Suppose a non-reflexive set A is infinite. Then select any member a of A and form the set A_1 of all elements of A except a. Select any member a_1 of A_1 and proceed to form the set A_2 excluding a_1. Select any member a_2 of A_2 and continue the process. A countably infinite set $B=\{a_1, a_2, a_3, \ldots\}$ emerges as a proper subset of A, proper since (at least) the element a is not included. So $A=B+C$, where C comprises all elements of A not in B. Now eliminate the element a_1 both from A (to get A') and from B to get $B'=\{a_2, a_3, \ldots\}$. Then $A'=B'+C$. There is a one-one correspondence between B and B', i.e. a_1 with a_2, a_2 with a_3, ... This can be extended by including the elements of C on both sides of the correspondence to give a one-one correspondence between A and A'.

Since A' is a proper subset of A, it follows that A is reflexive. But A is taken at the outset as non-reflexive. Hence, the non-reflexive set cannot be infinite; it must be finite as stated in the converse theorem.

The proof, however, is defective. A *finite* sequence $a_1, a_2, a_3, \ldots$ can be selected from A. But can an *infinite* sequence be selected? To say that it can is, as yet, no more than an intuitive extension from the finite to the infinite. Can it be justified? Mathematicians are agreed that it cannot be justified, and that the only way out of the difficulty is to impose it as an axiom in the theory of sets. This is the *Axiom of Choice* which guarantees that an infinite set A contains within it a countably infinite sequence of elements $a_1, a_2, a_3, \ldots$. But agreement goes no further. Mathematicians have been and are still divided on the question, not whether the Axiom is 'true' (which is not the question to ask about an axiom), but whether it should be permitted to take its place with the other axioms in the theory of sets. If the Axiom is so permitted — and at least it is known to be consistent with the other axioms — then the reflexive property of infinite sets follows, as do many other results in transfinite arithmetic and elsewhere in mathematics.† If the Axiom is not permitted, then all the results which depend on it must also be disallowed. Here is an awkward choice for mathematicians to make; most accept the Axiom more or less reluctantly while there are some who refuse to admit it. Even in mathematics all is not agreed.

4.9. Exercises

1. At a particular moment of time, take U as the set of all people ever born in the world. Consider the following sets

A = present population of the world,
B = all ancestors of the present population of the world,
C = present population of the United Kingdom (Great Britain and N. Ireland),
D = present population of Ireland (Republic of Ireland and N. Ireland).

Show that $B' = A$ and that $C \subset A$, $D \subset A$. What is the set $C \cup D$? Show that the present population of Great Britain is the set $C \cap D'$, of the Republic of Ireland the set $C' \cap D$ and of N. Ireland the set $C \cap D$.

2. Two dice are thrown, giving digits n_1 and n_2. List the set A of all possibilities, where $A = \{(n_1, n_2) \mid n_1 \epsilon J_6, n_2 \epsilon J_6\}$ for $J_6 = \{1, 2, 3, 4, 5, 6\}$. How many elements has A? How many elements are there in each subset A_r where A_r consists of the set of throws with sum $n_1 + n_2 = r$ $(r = 2, 3, \ldots 12)$?

† For example, the result that any two cardinal numbers (d, c and others) are commensurable, one of the relations $>$, $=$ and $<$ holding between them. There are also critical results in the calculus which depend on the Axiom.

3. In Ex. 2, interpret the set $B=\{n \mid n_1\epsilon J_6, n_2\epsilon J_6, n=n_1+n_2\}$ and show that $B=\{2, 3, \dots 12\}$.

4. The high table at a banquet has a row of seats for the chairman, two V.I.P.'s (A_1 and A_2) and their wives (B_1 and B_2). Show that there are 24 elements in the set of all seating arrangements, given only that the chairman is in the centre. How many elements are there in the subset when (apart from the chairman) the sexes alternate and in that when B_1 and B_2, bitter rivals, are separated by more than one person?

5. Indicate the sets $A\cap B$, $A\cap B'$, $A'\cap B$ and $A'\cap B'$ on a Venn Diagram, showing that they are disjoint and partition the universal set U.

6. Draw a Venn Diagram for $A'\cup B$ and show that $(A'\cup B)'=A\cap B'$. Translate into the form of the rules of 4.3 and check by rules 9 and 10.

7. Use Venn Diagrams to show: $A\cap(B\cup C)=(A\cap B)\cup(A\cap C)$; which rule of 4.3 is this?

8. Draw Venn Diagrams to show: $A\cup(B\cup C)=(A\cup B)\cup C$; $A\cup A=A$; $A\cup A'=U$. Identify as rules of 4.3. Write and establish the corresponding rules for $\cap$.

9. By showing that each is A, establish: $(A\cup B)\cap A=(A\cap B)\cup(A\cap B')$.

10. An interviewer asks: Do you like this chocolate bar? He conducts interviews and records the results:

	Yes	*No*	*Don't know*
Men	10	20	5
Women	20	15	5
Children	10	5	10

Write A = set of adults, C = set of women and children, Y = set of 'yes' answers and N = set of 'no' answers. Identify and find the number of each of the sets: A', $A\cap C$, $(Y\cup N)'$, $A\cap(Y\cup N)'$. Noting that the 10 yes-men make up the set $C'\cap Y$, express each of the above categories in this form.

11. *A Boolean algebra of two elements.* The set $\{0, 1\}$ is closed under + and ×, as defined by the tables of 4.4. Check that $0+(1+0)=(0+1)+0$, $0+1=1+0$, $1+1=1$, $1\times(0+1)=1\times 0+1\times 1$ and $1+(0\times 1)=(1+0)\times(1+1)$. Complete the checking of the operational rules of 4.3.

12. *Difference of sets.* If $A-B$ denotes AB' for $B\subseteq A$, show that $U-A=A'$, $A-A=0$ and $A-0=A$. Show that $(A-B)-C=A-(B+C)$ can be written if B and C are disjoint and such that $B\subset A$ and $C\subset A$.

13. Establish a one-one correspondence for $\{2n \mid n\epsilon J_5\}\sim\{2n-1 \mid n\epsilon J_5\}$ where $J_5=\{1, 2, 3, 4, 5\}$. Deduce that there are as many even as odd integers from 1 to 10 inclusive. Why is this not true for 1 to 9 inclusive? Generalise.

14. The results of Ex. 10 can be expressed: if P is the set of adults who say yes and Q the set of women and children who say yes, then $n(P)=30$, $n(Q)=30$, $n(PQ)=20$. Check that $n(P+Q)=n(P)+n(Q)-n(PQ)$. Interpret $n(P+Q)'=60$ as the number of those who say no or don't know.

15. An insurance company classifies a set of 50 'lives' according as they are men (or women), married (or single), British (or foreign) and gets the overlapping groups: Men 18, Married 20, British 39; Married men 7, British men 14, British married persons 16; Married and British men 6. Analyse into non-overlapping groups, using the method of 4.6. Show that, out of 11 foreign 'lives', 4 are single women.

16. Because of an error of transcription, the number of married 'lives' in Ex. 15 is recorded as 25 (not 20). Check that the data are then inconsistent.

17. *Removal of elements from a countable set.* From a countable set A, a set B is got by removing a countable number of elements. Show that B is countable. Interpret when A is finite. If A is countably infinite, show that B must be countably infinite when a finite number of elements is removed, and use the set of positive integers to show that B may be finite or countably infinite when a countably infinite number of elements is removed.

18. If a and b are any rationals, arrange the elements $a+b\sqrt{2}$ in double array, order by diagonals and show that $R(\sqrt{2})$ is countably infinite.

19. If $y=x/(1-x^2)$ is defined on the domain of real numbers $0<x<1$ show that the range of y is the set of all positive real numbers. Proceed: given any real $y>0$, show that there is a corresponding

$$x=\sqrt{\{(1/4y^2)+1\}}-(1/2y)>0.$$

Then, by showing that $\sqrt{\{(1/4y^2)+1\}}<(1/2y)+1$, deduce that $x<1$. Hence establish a one-one correspondence between the set

$$X=\{x \mid x \text{ a real number, } 0<x<1\}$$

and the set of all real positive numbers.

20. Extend the result of Ex. 19 by defining y on the domain $-1<x<1$ and establishing a one-one correspondence between the set of real numbers $-1<x<1$ and the set of all real numbers.

*21. Show that the set of Gaussian integers (3.2) is countable.

*22. Consider the set S of all roots of all polynomial equations with rational coefficients. Show that only the corresponding polynomials with integral coefficients need be taken, a countably infinite set (4.7). Put the roots in sequence by considering the n roots of each polynomial equation of degree n and deduce that S is countably infinite (even if duplication is not eliminated).

23. Show that the 'rule' that the part is smaller than the whole is equivalent to: all sets handled are non-reflexive. Can you say that the 'rule' holds only for finite sets, failing for any infinite set?

CHAPTER 5

STATEMENTS AND PROBABILITY

5.1. Statements. A *simple statement* is an assertion. In a given case, it is either true or false; it cannot be both. Statements can be made without regard to their truth, for they may be sometimes true and at other times false. As a notation, represent simple statements by small letters: p, q, r, ... For example:

(i) p: the child is a boy; q: the child is tall; r: the child has fair hair

(ii) p: equity prices are high; q: all equity prices are rising; r: some equity prices are rising

(iii) p: the triangle $\triangle$ is equilateral; q: the triangle $\triangle$ is isosceles; r: the triangle $\triangle$ has at least two unequal angles.

It is clear, from these examples, that the entities and features referred to must be understood, e.g. 'child', 'equity prices', 'equilateral'. In some cases, moreover, definitions need to be supplied, as in the use of 'tall' which might be taken as over 65 inches for a twelve-year-old child, and for 'high' which could be specified, for equities on Wall Street, as Standard and Poor's index (for 420 industrials) of over 200, the level of 1947–9 being 100.

A *compound statement* is formed from simple ones by the use of defined *connectives*. Three usual connectives correspond to 'not', to 'or' and to 'and':

Negation	$\sim p$: not p	Assertion of the negation of p.
Disjunction	$p \vee q$: p or q	Assertion of either p or q (or both).
Conjunction	$p \wedge q$: p and q	Assertion of p and q together.

Two other essential connectives are conditions, leading later to assertions of implication and equivalence:

Conditional	$p \rightarrow q$: if p then q.
Bi-conditional	$p \leftrightarrow q$: if p then q and if q then p.

No causal relationship is intended here. Like a simple statement, any compound statement such as $p \rightarrow q$ may be true or false (but not both) under given circumstances. Several connectives may be used in making further compound statements from those already constructed. Various examples illustrate:

(i) $\sim p$: the child is a girl
$q \wedge r$: the child is tall and fair-haired
$(\sim p) \rightarrow q$: if the child is a girl, then she is tall
(ii) $p \wedge q$: equity prices are high and rising
$\sim r$: no equity price is rising
$(\sim q) \wedge r$: some but not all equity prices are rising
$q \rightarrow r$: if all prices are rising, then some prices are rising
(iii) $p \vee q$: the triangle Δ is either isosceles or equilateral
$\sim r$: the triangle Δ has no unequal angles
$p \leftrightarrow (\sim r)$: if the triangle Δ is equilateral then it has no unequal angles, and conversely.

In a given problem, there will be a certain number of logical possibilities, perhaps only a few, more probably a large number. It will be enough to illustrate with a simple example where the logical possibilities are few. Consider a triangle Δ with reference to equal or unequal angles. There are five cases only in the set of all logical possibilities, as shown below in relation to whether the statements p, q and r of (iii) above are true (T) or false (F):

		p	q	r
Case	*Angles A, B, C*	*Equilateral*	*Isosceles*	*At least* 2 *unequal angles*
1	$A \neq B \neq C$	F	F	T
2*a*	$A \neq B = C$	F	T	T
2*b*	$B \neq C = A$	F	T	T
2*c*	$C \neq A = B$	F	T	T
3	$A = B = C$	T	T	F

However, as far as the three statements are concerned (though not necessarily other statements), there are only three different logical possibilities, since those labelled 2*a*, 2*b* and 2*c* all have p false, q and r true. It is enough to carry only three cases: 1 — all angles unequal; 2 — two angles equal and the other unequal; 3 — all angles equal.

Statements such as p, q and r are true in some cases and false in the others, and so are various compound statements, as shown in the table:

Case	p	q	r	$q \wedge r$	$p \vee r$
1	F	F	T	F	T
2	F	T	T	T	T
3	T	T	F	F	T

The interpretation of $q \wedge r$ is: the triangle Δ is isosceles and has at least two unequal angles, which is true in case 2 and false in cases 1 and 3. However, $p \vee r$ is true in all cases, since it means: the triangle Δ is equilateral or it has at least two unequal angles.

Most of our interest in statements is concentrated on those which are true in all logical possibilities, or false in all. One example is shown in the above table: $p \vee r$ is true in all three possible cases. Such a statement is called *logically true*. If a statement is false in all cases (as the statement $p \wedge r$), then it is called *logically false*.

The conditional connectives → and ↔ are most important when logically true or false. Consider two examples:

Case	p	q	$p \rightarrow q$
1	F	F	(T)
2	F	T	(T)
3	T	T	T

The statement $p \rightarrow q$*:* if p then q. As the table shows, $p \rightarrow q$ is true in case 3. There is an uncertainty in the other two cases where p is false; the statement $p \rightarrow q$ would not seem to apply. But the statement is not false and so (by convention at least) it can be marked as true, as shown by (T) in the table. Accepting this situation, we say that $p \rightarrow q$ is true in all cases, i.e. logically true. The only situation which would make $p \rightarrow q$ false (p true but q false) does not arise. The interpretation is that, in all cases, if Δ is equilateral, then it is isosceles. This situation is described as: *p implies q*.

Case	p	$\sim r$	$p \leftrightarrow (\sim r)$
1	F	F	T
2	F	F	T
3	T	T	T

The statement $p \leftrightarrow (\sim r)$*:* if p then $(\sim r)$ and if $(\sim r)$ then p. The table shows that the statements p and $(\sim r)$ are true (false) in precisely the same cases. Hence, $p \leftrightarrow (\sim r)$ is logically true. It could be falsified if p true but q false, or if p false but q true. Neither situation arises. The interpretation is that, in all cases, if △ is equilateral, then it has no unequal angles, and conversely. The situation is described as: p and $(\sim r)$ are *equivalent.*

The following definitions are thus to be made for any statements:

DEFINITION: p **implies** q *if* $p \to q$ *is logically true, i.e. 'if p then q' is true in all logical possibilities; p and q are* **equivalent** *if* $p \leftrightarrow q$ *is logically true, i.e. 'if p then q' and 'if q then p' are true in all logical possibilities.*

Notice that p implies q rules out only one situation (p true but q false); p and q equivalent rules out two situations (p true but q false, p false but q true).

Two logically false statements of interest can be illustrated, again with reference to statements on the triangle △.

Case	p	r	$p \leftrightarrow r$
1	F	T	F
2	F	T	F
3	T	F	F

The statement $p \leftrightarrow r$*:* if p then r and if r then p. The table shows that the statement is logically false. The situations ruled out are those in which p and r are both true or both false. Here p and r are *contradictories.* It reflects, of course, the equivalence of p and $(\sim r)$; the statement that △ is equilateral contradicts the statement that △ has at least two unequal angles.

Case	p	$q \wedge r$	$p \wedge (q \wedge r)$
1	F	F	F
2	F	T	F
3	T	F	F

The statement $p \wedge (q \wedge r)$*:* p and q and r together asserted. Again the statement is logically false, but now the only situation ruled out is p true and $(q \wedge r)$ true. These two (simpler) statements cannot both be true; they are described as *contraries.* In this example, p is that △

is equilateral; $(q \wedge r)$ is that △ is both isosceles and has one angle unequal to the other two. There is no triangle in which both these things are true.

The tables with T and F entries against the various logical possibilities are instances of *truth tables*; they are useful tools in analysing the logic of statements. Of even more use, however, is the representation of statements in terms of sets, illustrated by means of Venn Diagrams.

5.2. Statements and sets. Consider sets of logical possibilities in a given problem. A statement p is specified; it is true in some and false in the other logical possibilities. Denote by P the set of logical possibilities for which p is true, the *truth set* of the statement p. Two bounds can be defined: the universal set U of all logical possibilities and the empty set ϕ. U is the truth set of a statement which is logically true, ϕ of a statement which is logically false:

$$\phi \subseteq P \subseteq U \quad \text{for any } P.$$

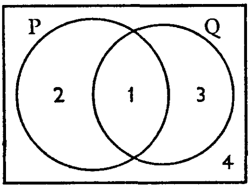


FIG. 5.2

If two statements p and q are specified, write P for the truth set of p and Q for the truth set of q. The sets P and Q serve to divide U into four disjoint and exhaustive sets (one or more of which may be empty). This is shown in the Venn Diagram of Fig. 5.2, the sets being numbered:

	1	2	3	4
Set in which:	$P \cap Q$	$P \cap Q'$	$P' \cap Q$	$P' \cap Q'$
p	T	T	F	F
q	T	F	T	F

Further: $P \cup Q$ = union of sets 1, 2 and 3 in which either p or q true,

$$(P \cup Q)' = P' \cap Q' = \text{set 4 in which both } p \text{ and } q \text{ false.}$$

Consequently, compound statements under the three connectives:

$$\sim = \text{'not'}; \quad \vee = \text{'or'}; \quad \wedge = \text{'and'}$$

are represented by sets according to the scheme:

Statement	*Truth Set*	*No. in Venn Diagram*
$\sim p$	P'	3 and 4
$p \vee q$	$P \cup Q$	1, 2 and 3
$p \wedge q$	$P \cap Q$	1

The connective $\sim$ corresponds to complement, $\vee$ to union $\cup$ and $\wedge$ to intersection $\cap$ in the operations on sets. The parallel is perfect so that:

THEOREM: *The algebra of statements under the connectives 'not', 'or', 'and' is a Boolean Algebra.*

Statements are to be translated into their truth sets and the operational rules of 4.3 applied. In this Boolean Algebra, any logically true statement plays the role of the upper bound U and any logically false statement the lower bound ϕ. In between are statements true in some and false in other logical possibilities.

The translation of implication and equivalence into set terms can be easily achieved. The statement $p \rightarrow q$ is false only when p is true but q false. If P and Q are the truth sets, $p \rightarrow q$ is false only in the set $P \cap Q'$, which is numbered 2 in the Venn Diagram. Hence the truth set of $p \rightarrow q$, comprising the union of sets 1, 3 and 4 in the diagram, is the complement of $P \cap Q'$:

$$\text{Truth set of } \; p \rightarrow q = (P \cap Q')' = P' \cup (Q')' = P' \cup Q$$

by Boolean Algebra. As a check, since P' is sets 3 and 4 and Q is sets 1 and 3 in the diagram, $P' \cup Q$ is sets 1, 3 and 4. It also follows that the statement $p \rightarrow q$ is equivalent to $(\sim p) \vee q$, each having the truth set $P' \cup Q$.

The statement p implies q is defined as $p \rightarrow q$ logically true. This means, in terms of sets:

$$P' \cup Q = U \quad \text{or} \quad P \cap Q' = \phi$$

i.e. the truth set of $p \rightarrow q$ is the universal set, the set in which $p \rightarrow q$ is false is the empty set. Check by Boolean Algebra:

If $P' \cup Q = U$, then $\phi = U' = (P' \cup Q)' = (P')' \cap Q' = P \cap Q'$.
Now if $P \cap Q'$ is empty, then P and Q' are disjoint and P must be contained in Q: $P \subseteq Q$. In terms of truth sets, 'p implies q' means $P \subseteq Q$. Equivalence follows immediately. The statements p and q are equivalent if p implies q and q implies p. In set terms, $P \subseteq Q$ and

$Q \subseteq P$, which means $P = Q$. The truth sets of p and q are the same. Hence:

THEOREM: *The statement p* **implies** *the statement q if and only if the truth set of p is a subset of the truth set of q* $(P \subseteq Q)$. *The statements p and q are* **equivalent** *if and only if the truth sets coincide* $(P = Q)$.

It also follows that, if p implies q without q implies p, then the truth set of p is a proper subset of the truth set of q $(P \subset Q)$.

An immediate consequence is: p implies q means $P \subseteq Q$ and this in its turn means $Q' \subseteq P'$, i.e. that $(\sim q)$ implies $(\sim p)$. Conversely, $(\sim q)$ implies $(\sim p)$ means $Q' \subseteq P'$, which means $P \subseteq Q$ or p implies q. Hence:

THEOREM: *The statements 'p implies q' and '$(\sim q)$ implies $(\sim p)$' are equivalent.*

This is a result which has a bearing on the methods of proof used in mathematics. If we are given a result or condition p and if we wish to deduce a consequential result or condition q, the direct method of proof is to show p implies q, i.e. if p then q in all logical possibilities. It is equally valid to use an indirect method of proof, to show $(\sim q)$ implies $(\sim p)$. That is, given p, assume the contrary $(\sim q)$ of what is to be established and go on to deduce $(\sim p)$. Hence a contradiction, both p and $(\sim p)$. This means that the contrary assumption $(\sim q)$ must be abandoned and q is established. This is the basis of the proof by *reductio ad absurdum*. As a simple illustration, take the following not-quite-trivial example. Given that the product of two even positive integers is even, it follows that n^2 odd implies n odd (n a positive integer). The proof is indirect. Take n^2 odd (given) and assume n even. Then $n^2 = n \times n$ is even. There is a contradiction, n^2 odd and n^2 even. Hence n is odd.

5.3. Necessary and sufficient conditions. In many mathematical developments, instead of proceeding from proposition to proposition (p implies q, q implies r, ...), we may take a property q and seek the conditions p for it to hold. In such a situation, we try to find p so that 'q if p', or 'q only if p', or 'q if and only if p'. A good deal of uncertainty and confusion of thought is possible in this apparently simple procedure. And not without reason. For one thing, the use of the alternatives, p implies q and $(\sim q)$ implies $(\sim p)$, serves to disguise the fact that these are the same. Confusion is then made worse by the common habit of employing various terms or modes of expression for the same thing. Since there *are* alternative terms in use, we can do no more than be on the lookout for the various disguises.

First, take p implies q. This can be expressed: if p then q in all logical possibilities; or, omitting the qualification, just: 'if p then q'. This can be turned around to read: 'q if p'. This is what is described as a *sufficient condition*; here p is a sufficient condition for q.

Second, take $(\sim q)$ implies $(\sim p)$. Omitting the same qualification, we can say: 'if $(\sim q)$ then $(\sim p)$', which is the same as: 'only if q, then p'. This is turned around to read: 'p only if q'. This is what is described as a *necessary condition*; here q is a necessary condition for p.

Confusion can be avoided only by recognising and keeping always in mind the simple fact that all the statements written in the two preceding paragraphs *are precisely the same.* If p and q have truth sets P and Q, all statements correspond quite simply to $P \subseteq Q$. The first run of statements stems from p implies q, the second from $(\sim q)$ implies $(\sim p)$. These two are equivalent; both mean $P \subseteq Q$. Hence, if p and q are such that $P \subseteq Q$ for their truth sets, then all the following statements and notations follow:

Implication	*One phrasing*	*Alternative phrasing*	*Terminology*
p implies q	if p then q	q if p	p is a sufficient condition for q
$\sim q$ implies $\sim p$	if $\sim q$ then $\sim p$	p only if q	q is a necessary condition for p

A similar table can be written for all the ways of writing: q implies p. It differs only in that p and q are interchanged. For example, if we are seeking p as a necessary condition for q, we would look at this second table, or (what is the same thing) interchange p and q in the table above.

Turn now from implication to equivalence, i.e. take *both* p implies q *and* q implies p. There is a completely symmetrical relationship between p and q. Setting down in full all the ways of putting the position: if p and q are such that their truth sets are equal $(P=Q)$ then all the following hold:

Equivalence	*One phrasing*	*Alternative phrasing*	*Terminology*
p and q equivalent	if p then q *and* if $\sim p$ then $\sim q$	q if and only if p	p is a necessary and sufficient condition for q
q and p equivalent	if q then p *and* if $\sim q$ then $\sim p$	p if and only if q	q is a necessary and sufficient condition for p

Notice that, in the phrase 'if and only if', the necessary part is 'only if' and the sufficient part is 'if'. A *necessary and sufficient condition* can only be established in two stages: *first* q only if p, meaning that for q to hold we must exclude all things other than what follows the 'only if' (i.e. exclude everything except p); *second* q if p, meaning that q holds if we include everything that follows the 'if' (i.e. include p); and so *together* q if and only if p, the necessary and sufficient condition for q.

Two examples from elementary geometry illustrate:

(i) Conditions are sought for q: $\triangle$ is an isosceles triangle. Write p: $\triangle$ is an equilateral triangle. Then p implies q. For, by elementary geometry, an equilateral triangle is isosceles. Hence q if p, and p is a *sufficient* condition for q. We have, in fact, overdone the conditions for an isosceles triangle. Certainly an equilateral triangle is isosceles, but many other triangles are isosceles too. p is not a necessary condition.

Turn things around and seek conditions for p: $\triangle$ is an equilateral triangle. Write q: $\triangle$ is an isosceles triangle. Then $(\sim q)$ implies $(\sim p)$. For, if $\triangle$ is not isosceles, it is not equilateral either. Hence p only if q, and q is a *necessary* condition for p. We have, in fact, not got enough to ensure that the triangle is equilateral. q is not a sufficient condition.

These two lines of reasoning are precisely the same. We put it one way or the other according to our view: seeking conditions for q (isosceles triangle) or seeking conditions for p (equilateral triangle).

(ii) Conditions are sought for q: the quadrilateral Q is a rectangle. More specifically, we seek conditions p in terms of the diagonals of Q.

Write p: the diagonals are equal and bisect at right angles. Then p implies q. This is indicated by A in Fig. 5.3 where the solid lines show the diagonals under p, and the dotted lines show the resulting quad-

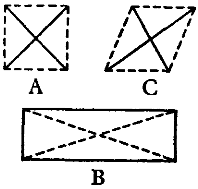


Fig. 5.3

rilateral Q (clearly a rectangle). Hence q if p, and p is a *sufficient condition* for q. Again, we have overdone it; Q is a rectangle all right, in fact a square.

Write p: the diagonals bisect each other. Then q implies p. This is indicated by B in the diagram, solid lines showing the rectangle (q) and dotted lines the diagonals under p. Hence q only if p, and p is a *necessary condition* for q. The position can also be put: $(\sim p)$ implies $(\sim q)$, i.e. if the diagonals do not bisect, then Q is not a rectangle. Here we do not have enough to ensure a rectangle; bisecting diagonals can produce a parallelogram (C of the diagram).

Write p: the diagonals are equal and bisect each other. Then p implies q, i.e. q if p. If we draw equal and bisecting diagonals, we must get a rectangle. Further, q implies p, or $(\sim p)$ implies $(\sim q)$, i.e. q only if p. We cannot draw a rectangle without diagonals equal and bisecting. So: q if and only if p, and p is the *necessary and sufficient condition* for q.

5.4. Probability. The concept of probability has to do with statements, with assertions or propositions. Consider a group of twelve-year-old boys; the statement: the boy is tall (over 65 inches) may be true for some and false for others in the group. Similarly, if we measure equity prices by Standard and Poor's industrials index, the statement: equity prices are rising is sometimes true and sometimes false. In each case, there are two alternatives: boys tall or not tall, prices rising or not rising. How 'likely' is such a statement to be true? Can we put a measure on the 'likelihood'? For example, we may say that 'not many' boys are tall and we may assess the 'chance' at well under $\frac{1}{2}$, say 10 : 90 or 10 chances out of 100. In considerations of this kind, a variety of rather ill-defined terms tends to crop up: probable, likely, chances for or against, degrees of confidence. All of them apply to a particular statement which is under discussion.

The same general ideas are often expressed in another way. A situation is considered in which the outcome is 'uncertain'. For example, some observations may be made on the movement of equity prices during a week. In a set of observations there is a range of 'outcomes', e.g. prices rise or fall by various amounts on different days. In one observation there is only one 'outcome'. If we do not know it, what is the chance that it is this or that? For the equity market on the current day, what is the chance of a price rise? Recent observations may assist us, but we may be reduced to saying that a rise and a fall are 'equally likely'; we assess the uncertain outcome at a chance of $\frac{1}{2}$ for a rise. Other rather ill-defined terms tend to appear: doubt, uncertainty, need for confirmation. They apply to outcomes in a particular situation.

The second formulation places the emphasis on the uncertainty of events, rather than on a statement which may or may not be true. But, however we look at the matter, the basic concept is a statement. We state a proposition: equity prices will rise; the chances are then assessed that the proposition is true.

To fix ideas consider two simple examples:

(i) This is a familiar, if rather artificial, kind of experiment for illustrating probabilities: two dice are thrown and the sum $n = n_1 + n_2$ of the two digits appearing is written down. We look for 7 or 11. How likely is the proposition that we get $n = 7$ or 11? The first task is to specify the set of all possible outcomes. If we look at the digits n_1 and n_2 separately, there are 36 pairs:

Case	n_1	n_2	n	*Case*	n_1	n_2	n	*Case*	n_1	n_2	n	*Case*	n_1	n_2	n
1	1	1	2	10	2	4	6	19	4	1	5	28	5	4	9
2	1	2	3	11	2	5	7	20	4	2	6	29	5	5	10
3	1	3	4	12	2	6	8	21	4	3	7	30	5	6	11
4	1	4	5	13	3	1	4	22	4	4	8	31	6	1	7
5	1	5	6	14	3	2	5	23	4	5	9	32	6	2	8
6	1	6	7	15	3	3	6	24	4	6	10	33	6	3	9
7	2	1	3	16	3	4	7	25	5	1	6	34	6	4	10
8	2	2	4	17	3	5	8	26	5	2	7	35	6	5	11
9	2	3	5	18	3	6	9	27	5	3	8	36	6	6	12

We may decide that the dice are without bias in the sense that all 36 outcomes are 'equally likely'. Six of them produce $n = 7$ (cases 6, 11, 16, 21, 26, 31) and two of them give $n = 11$ (cases 30, 35). Together, the proposition $n = 7$ or 11 is true in 8 cases out of 36. It is then an easy step to assess the chance that the proposition is true at 8 : 28 or $8/36 = 2/9$.

To illustrate, as in the example of the triangle in 5.1, that the mode of specification of the logical possibilities is not unique, define outcomes solely in terms of the sum n obtained at a throw. There are 11 of them:

Sum n	2	3	4	5	6	7	8	9	10	11	12
Chance μ (out of 36)	1	2	3	4	5	6	5	4	3	2	1

Even if there is no bias in the dice, we decide that the outcomes are not 'equally likely' and we attach assessments of the relative likelihood of one outcome as opposed to another. One set of assessments μ, adding to 36, is shown above. This is, in fact, obtained by going through the same process as before (e.g. $n=7$ arises in 6 out of 36 cases); but it may be got by some other procedure (e.g. by experiment). The point is that something must be assumed: equal chances for the 36 outcomes, unequal chances as shown for the 11 outcomes.

(ii) The heights of a group of 100 twelve-year-old boys are recorded. We are interested in tall boys (over 65 inches). How likely is the proposition that one selected boy is tall? The specification of all possible outcomes raises difficulties. In a narrow sense, there are 100, the actual records. But this is not the problem really considered; the records are a 'sample' in some sense from a wider range of possible outcomes. Suppose we can set limits to the height x inches, e.g. $50 \leqslant x \leqslant 70$. Conceptually, x is any real number in this range, a noncountable infinity of outcomes. In practice, x is certainly a rational number, usually one with 10, 100, 1000, ... in the numerator according to the measuring equipment used, e.g. 100 if heights are read to the second decimal place. Then $x=n/100$ and, for $50 \leqslant x \leqslant 70$, n is any integer from 5000 to 7000 inclusive, i.e. x can take a finite but large (here 2001) number of values. There are too many to handle and some rounding is needed in specifying the outcomes, as in the following approximation:

Height x	52	54	56	58	60	62	64	66	68	70
Chance μ (out of 100)	1	1	3	11	26	32	16	7	2	1

Here height is specified in ranges of 2 inches and the interpretation (e.g.) of the first case is: 52 ± 1, i.e. over 51 but not over 53. The question of assessing the chance of getting each x (here reduced to

10 cases) still arises. This may be done in various ways, from theoretical considerations or from empirical distributions of heights. The set of assessments μ in the table is illustrative; it gives the chance of a tall boy as 1/10, i.e. $7+2+1=10$ chances out of 100.

Probability theory is an attempt to *quantify* the idea of the chance of a statement being true and to devise an *algebra* to handle the chances. The problem is of a kind lending itself to a variety of treatments and, indeed, to controversy. One approach is suggested here: to switch from properties of statements to the corresponding properties of their truth sets. The relation between them is given in 5.2. If the statement a_1 has truth set A_1 and if a_2 has truth set A_2, then:

Statement	$\sim a_1$	$a_1 \vee a_2$	$a_1 \wedge a_2$	$a_1 \rightarrow a_2$
Truth set	A_1'	$A_1 \cup A_2$	$A_1 \cap A_2$	$A_1' \cup A_2$

What we must do is to attach a *measure* $\mu(A)$ to a set A, and then to equate to the *probability* $P(a)$ of the statement a with truth set A.

5.5. Probability measure. The definition of the set measure $\mu(A)=P(a)$ must be such that certain requirements are satisfied, those set by the everyday use of probability. The chance of throwing a 5 in one throw of a die is $\frac{1}{6}$, if the die has no bias; the chance of a 6 is the same. What is the chance of getting a 5 *or* a 6 in one throw? The answer we expect is $\frac{1}{6}+\frac{1}{6}=\frac{1}{3}$. What is the chance of getting a 6 *followed* by a 6 in two throws? We expect $\frac{1}{6}\times\frac{1}{6}=\frac{1}{36}$. Similarly, if we ignore the fact that bookies work for profit, we interpret odds of 3 to 1 against a horse in one race as: chance of horse winning is $\frac{1}{4}$. For a horse in another race, odds of 5 to 3 against represent a chance of winning of $\frac{3}{8}$. What is the chance that either one horse wins, that both horses win? We expect $\frac{1}{4}+\frac{3}{8}=\frac{5}{8}$ (or 5 to 3 on) for either; $\frac{1}{4}\times\frac{3}{8}=\frac{3}{32}$ (29 to 3 against) for both.

Such requirements can be translated into properties of the probability $P(a)$ of a statement a, matched by properties of the measure $\mu(A)$ of the corresponding truth set. With a little care, and leaving only one matter open, we state the object of the exercise of defining $\mu(A)=P(A)$ in the following terms:

Probability measure $P(a)$	*Set measure* $\mu(A)$
(i) $P(a)=0$ if and only if a is logically false $P(a)=1$ if and only if a is logically true $0 \leqslant P(a) \leqslant 1$ any a	$\mu(A)=0$ if and only if $A=\phi$ $\mu(A)=1$ if and only if $A=U$ $0 \leqslant \mu(A) \leqslant 1$ any A, $\phi \subseteq A \subseteq U$
(ii) $P(a_1 \vee a_2)=P(a_1)+P(a_2)$ if a_1 and a_2 are exclusive ($a_1 \wedge a_2$ logically false)	$\mu(A_1 \cup A_2)=\mu(A_1)+\mu(A_2)$ if $A_1 \cap A_2=\phi$ (A_1 and A_2 disjoint)
(iii) $P(a_1 \wedge {}_2)=P(a_1) \times P(a_2)$ if a_1 and a_2 are independent (in some way to be defined)	$\mu(A_1 \cap A_2)=\mu(A_1) \times \mu(A_2)$ for certain types of sets A_1 and A_2

Requirement (i) is that a probability lies between 0 (logically false) and 1 (logically true). Requirement (ii) is the additive or disjunctive property of exclusive or non-overlapping statements, i.e. the chance of *either* is the sum of the separate chances. Requirement (iii) is the multiplicative or conjunctive property for statements which do not depend on each other (in a way to be made precise), i.e. the chance of *both* is the product of the separate chances.

The question of a suitable definition of $\mu(A)=P(a)$ is pursued here only in a simple case: the set U of all logical possibilities (outcomes) is finite. The reason for the simplicity is easily given in general terms. If there is only a finite number n of discrete outcomes, then we *can* if we wish say that they are equally probable. In terms of sets, we can assert:

$\mu(A_r)=\mu$ (constant, all r) if A_r is the truth set of the rth outcome.

The n sets A_r ($r=1, 2, \ldots n$) have U as their union:

$$A_1 \cup A_1 \cup \ldots \cup A_n = U.$$

Then, by the additive property (ii), extended to several sets:

$$\mu(A_1)+\mu(A_2)+ \ldots +\mu(A_n)=\mu(U)=1.$$

So: $$n\mu=1 \quad \text{i.e. } \mu=\frac{1}{n}.$$

This makes good sense: if there are n equi-probable outcomes, the chance of one of them is $\frac{1}{n}$. It is illustrated by example (i) of 5.4.

The case we do not consider further arises when U is countably or non-countably infinite, i.e. an infinite number of possible outcomes. Any attempt to define set measure and probability in this case takes us too far into measure theory. That the case is by no means unimportant is seen by reference to example (ii) of 5.4. The universal set U is here of the form $(x_1 \leqslant x \leqslant x_2)$ where x inches is height (limits of x_1 and x_2 being set) and where x is a real number. The kind of probability measure sought is then $P(x > 65)$; the set for which $x > 65$ is non-countably infinite like U itself. This is the situation considered in theoretical statistics which deals with a 'random variable' x subject to a defined probability distribution.* The difficulty when U is infinite is that we can *not* assign a non-zero measure to each outcome if we say that they are equi-probable.

The procedure for defining probability measure in a finite set of outcomes draws upon the results of 4.6 for counting sets. If there are n outcomes, assign the *weight* μ_r to the rth outcome $(r = 1, 2, 3, \ldots n)$ in such a way that $\sum\limits_r \mu_r = 1$. This is a matter of appropriate assumption in each case. A statement a has a finite truth set A, a subset of the set U of all outcomes. Let the number of elements in A be $n(A)$ where $0 \leqslant n(A) \leqslant n$. Add the weights attached to the elements of A and define as the measure of A, i.e. $\mu(A) = \sum\limits_A \mu_r$ where $\sum\limits_A$ extends over the $n(A)$ elements of A. Finally, the probability $P(a)$ of the statement a is $\mu(A)$.

DEFINITION: *The* **measure** *of a subset A of a finite set of possible outcomes with weights $\mu_r (\sum\limits_r \mu_r = 1,\ r = 1, 2, 3, \ldots n)$ is $\mu(A) = \sum\limits_A \mu_r$ where $\sum\limits_A$ is summation over elements of A and the* **probability** *of the statement a with truth set A is $P(a) = \mu(A)$.*

Properties (i) and (ii) are then satisfied. If a is logically false, $A = \phi$ and $\mu(A) = 0$; if a is logically true, $A = U$ and $\mu(A) = \sum\limits_r \mu_r = 1$. For any A, $\sum\limits_A \mu_r \geqslant 0$ and $\sum\limits_A \mu_r \leqslant \sum\limits_r u_r = 1$. So $0 \leqslant \mu(A) \leqslant 1$, as required. Further, if A_1 and A_2 are disjoint, then (as in 4.6) the number of elements in A_1 and the number of elements in A_2 add to the number of elements in $A_1 \cup A_2$. So do the weights μ_r and

$$\mu(A_1 \cup A_2) = \mu(A_1) + \mu(A_2).$$

* See Goldberg: *Probability, An Introduction* (Prentice-Hall, 1959).

The interpretation of property (iii) is left over.

In the particular case of *equi-probable measure:*

$$\mu_r = \mu \quad (r = 1, 2, 3, \ldots n).$$

Then: $$1 = \sum_r \mu_r = \sum_r \mu = n\mu \quad \text{i.e. } \mu = \frac{1}{n}$$

and for a set A of $n(A)$ elements: $\mu(A) = \sum_A \mu = n(A)\mu = \dfrac{n(A)}{n}$. Hence:

THEOREM: *If n outcomes are equi-probable, then the probability of a statement a is $P(a) = \dfrac{n(A)}{n}$ where $n(A)$ is the number of outcomes for which a is true.*

As an illustration, consider example (i) of 5.4: throws of two dice, assumed to be without bias. There are 36 equi-probable outcomes and $\mu_r = 1/36$ (all r). For the statement a that the sum of the digits is 7 or 11, the truth set consists of 8 outcomes: $P(a) = 8/36 = 2/9$. On the other hand, specify 11 outcomes with μ_r $(r = 1, 2, \ldots 11)$ as given in 5.4. This is *not* an equi-probable specification. The statement a has truth set A: the 6th outcome ($\mu_r = 6/36$) and the 10th ($\mu_r = 2/36$). Hence $\mu(A) = 6/36 + 2/36 = 2/9$ and $P(a) = 2/9$ again.

5.6. Properties of probability measure. The statements a_1 and a_2 are *exclusive* if both cannot be true together, i.e. if $a_1 \wedge a_2$ is logically false and $P(a_1 \wedge a_2) = 0$. The corresponding truth sets are disjoint: $A_1 \cap A_2 = \phi$. This is the case already discussed, giving

$$P(a_1 \vee a_2) = P(a_1) + P(a_2).$$

However, if a_1 and a_2 are *not exclusive*, then $P(a_1 \wedge a_2) \neq 0$ and $A_1 \cap A_2 \neq \phi$, i.e. A_1 and A_2 overlap. The number of elements in $(A_1 \cup A_2)$ is then given by the result of 4.6, i.e.

$$n(A_1 \cup A_2) = n(A_1) + n(A_2) - n(A_1 \cap A_2)$$

and similarly for the weights μ_r so that

$$\mu(A_1 \cup A_2) = \mu(A_1) + \mu(A_2) - \mu(A_1 \cap A_2)$$

i.e. $$P(a_1 \vee a_2) = P(a_1) + P(a_2) - P(a_1 \wedge a_2) \quad \ldots\ldots\ldots\ldots\ldots(1)$$

Consider the statement: a_1 given a_2. As a statement, it has a probability — the probability of a_1 given a_2. There is a different statement: a_2 given a_1 and so a different probability — the probability of a_2 given a_1. So:

DEFINITION: *The* **conditional probability** $P(a_1 \mid a_2)$ *is the probability of a_1 given a_2; similarly $P(a_2 \mid a_1)$ is the probability of a_2 given a_1.*

The truth set of a_1 given a_2 is $A_1 \cap A_2$, but this must be related, *not* to the universal set U, but to the set A_2 since a_2 is given. Hence to get $P(a_1 \mid a_2)$ the measure $\mu(A_1 \cap A_2)$ must be related to $\mu(A_2)$. Strictly: a new universal set A_2 is taken and the weights of the elements in A_2 re-scaled to add to unity. The original weights are μ_r adding to $\sum_{A_2}\mu_r = \mu(A_2)$. They are re-scaled to $\frac{\mu_r}{\mu(A_2)}$ adding to $\sum_{A_2}\frac{u_r}{\mu(A_2)} = \frac{1}{\mu(A_2)}\sum_{A_2}\mu_r = 1$. $P(a_1 \mid a_2)$ is the (re-scaled) measure of $A_1 \cap A_2$:

$$P(a_1 \mid a_2) = \sum_{A_1A_2}\frac{\mu_r}{\mu(A_2)} = \frac{1}{\mu(A_2)}\sum_{A_1A_2}\mu_r = \frac{\mu(A_1 \cap A_2)}{\mu(A_2)}\,.$$

Similarly: $$P(a_2 \mid a_1) = \frac{\mu(A_1 \cap A_2)}{\mu(A_1)}\,.$$

Since $\mu(A_1) = P(a_1)$, $\mu(A_2) = P(a_2)$ and $\mu(A_1 \cap A_2) = P(a_1 \wedge a_2)$, we have:

$$P(a_1 \wedge a_2) = P(a_1 \mid a_2)P(a_2) = P(a_2 \mid a_1)P(a_1) \quad \ldots\ldots\ldots\ldots(2)$$

Suppose a_1 and a_2 are such that $P(a_1 \mid a_2) = P(a_1)$. Then, by (2):

$$P(a_2 \mid a_1) = \frac{P(a_1 \mid a_2)P(a_2)}{P(a_1)} = P(a_2)\,.$$

Hence, it is true *both* that $P(a_1 \mid a_2) = P(a_1)$ *and* that $P(a_2 \mid a_1) = P(a_2)$. In other words, each statement is irrelevant to the other. In this case, the statements a_1 and a_2 are said to be *independent*. Result (2) then gives:

$$P(a_1 \wedge a_2) = P(a_1)P(a_2) \quad \text{if } a_1 \text{ and } a_2 \text{ are independent} \ \ldots\ldots(3)$$

Just as exclusive statements allow straight addition of probabilities, so independent statements allow straight multiplication of probabilities.

An assembly of the results obtained, and specifically the results (1), (2) and (3) just established, gives the properties of probability measure:

THEOREM

(i) **Bounds.** $0 \leqslant P(a) \leqslant 1$ *for any statement a*
and $P(\sim a) = 1 - P(a)$

(ii) **Disjunction.** $P(a_1 \vee a_2) = P(a_1) + P(a_2) - P(a_1 \wedge a_2)$ *for any statements a_1 and a_2.*
In particular: $P(a_1 \vee a_2) = P(a_1) + P(a_2)$ *if a_1 and a_2 are* **exclusive.**

(iii) **Conjunction.** $P(a_1 \wedge a_2) = P(a_1 \mid a_2)P(a_2) = P(a_2 \mid a_1)P(a_1)$ *for any statements a_1 and a_2.*
In particular: $P(a_1 \wedge a_2) = P(a_1)P(a_2)$ *if a_1 and a_2 are* **independent.**

Note that $P(\sim a) = 1 - P(a)$ follows from the particular case of (ii) since $(\sim a)$ and a are exclusive and such that $\sim a \vee a$ is logically true. Hence:

$$P(\sim a) + P(a) = P(\sim a \vee a) = 1 \quad \text{i.e. } P(\sim a) = 1 - P(a).$$

The *additive* result (ii) means that the chance of *either* statement is the sum of the separate chances if and only if the statements are *exclusive*. Otherwise, the chance of both statements must be allowed for. The *multiplicative* result (iii) means that the chance of *both* statements is the product of the separate chances if and only if the statements are *independent*. Otherwise the conditional probability of one statement given the other is involved.

Property (ii) extends in an obvious way to the case of three or more statements. If they are not exclusive, the difficulty of allowing for overlap gets greater as more statements are taken together, as for counting sets in 4.6. If they are exclusive, the probabilities of the separate statements simply add. Moreover, statements $a_1, a_2, a_3, \ldots$ are *exclusive and exhaustive* if $a_i \wedge a_j$ is logically false $(i \neq j)$ and if $a_1 \vee a_2 \vee a_3 \ldots$ is logically true. Then:

$$P(a_1) + P(a_2) + P(a_3) + \ldots = 1.$$

Notice that, if $a_1, a_2, a_3 \ldots$ are *not exclusive* (whether they exhaust all outcomes or not), it is perfectly possible that:

$$P(a_1) + P(a_2) + P(a_3) + \ldots > 1 \quad \text{for some } a_1, a_2, a_3, \ldots.$$

Property (iii) can be developed to give a result of basic importance in the problem of inference. Consider the conjunction of a statement a_r with another statement a. Then (iii) gives:

$$P(a_r \mid a)P(a) = P(a \mid a_r)P(a_r). \quad \text{....................(4)}$$

Now suppose a set $a_1, a_2, a_3, \ldots$ of *exclusive and exhaustive* statements is specified, given a, so that $\sum_r P(a_r \mid a) = 1$. In other words, if a is known to be true, then *one* of $a_1, a_2, a_3 \ldots$ is true. The question is: which one? The result (4) can be written for each r $(r = 1, 2, 3, \ldots)$ and added:

$$P(a)\sum_r P(a_r \mid a) = \sum_r P(a \mid a_r)P(a_r) \quad \text{i.e. } P(a) = \sum_r P(a \mid a_r)P(a_r).$$

Write $\lambda = P(a)$, not depending on a_r. Then (4) becomes:

$$P(a_r \mid a) = \frac{1}{\lambda} P(a \mid a_r)P(a_r) \quad (r = 1, 2, 3, \ldots) \text{...........(5)}$$

The result (5), known as *Bayes' Theorem* after Bayes (*d.* 1761), states that the *posterior probability* $P(a_r \mid a)$ is proportional to the *prior probability* $P(a_r)$, multiplied by the chance of the known a on a_r, $P(a \mid a_r)$. For any one statement a_r, its chance with nothing known (i.e. the prior probability) is compared with its chance when the evidence a is known (i.e. the posterior probability). In passing from one to the other, the factor which comes in is $P(a \mid a_r)$, the chance of the evidence a on the basis of the particular a_r. This factor varies from one a_r to another. The idea is to select that a_r which has the *greatest posterior probability*, i.e. the greatest $P(a \mid a_r)P(a_r)$. For this, the prior probabilities $P(a_r)$ must be known.

So far, so good. Bayes' Theorem is not subject to criticism; it is an established property of conditional probabilities. The difficulty is that the prior probabilities $P(a_r)$ are not in fact known. A second step is now taken: assume all $P(a_r)$ are equal to a constant μ. This is *Bayes' Postulate*, i.e. when nothing is known about prior probabilities $P(a_r)$, assume them equal. Then (5) is:

$$P(a_r \mid a) = \frac{\mu}{\lambda} P(a \mid a_r).$$

So, to select the greatest $P(a_r \mid a)$, we select a_r with the greatest $P(a \mid a_r)$. On Bayes' Postulate, the selection of the particular a_r is solely on the basis of $P(a \mid a_r)$, the chance of getting the known a from a_r. This is one of the most controversial matters in the theory of inference: the selection of a_r from a set of possible outcomes, on the evidence of a.

5.7. Examples of probability measure. The simplest and most familiar examples of probability (finite number of outcomes) are taken from such experiments as dice-throwing and card-dealing. To illustrate exclusive and non-exclusive statements, examine the digit (1, 2, 3, 4, 5 or 6) shown in the throw of a die. The two statements:

$$a_1\text{: digit 1, 2 or 3} \quad a_2\text{: digit 4, 5 or 6}$$

have each the probability $\frac{1}{2}$ if the die has no bias. They are exclusive and exhaustive; the probabilities add: $\frac{1}{2}+\frac{1}{2}=1$. But consider the two statements:

$$b_1\text{: digit 1, 2, 3 or 4} \quad b_2\text{: digit 3, 4, 5 or 6.}$$

Each has the probability $\frac{2}{3}$, so that

$$P(b_1)+P(b_2)=\tfrac{2}{3}+\tfrac{2}{3}=\tfrac{4}{3}>1.$$

This is so because b_1 and b_2 are not exclusive. To get $P(b_1 \vee b_2)$, use (ii):

$$\begin{aligned} P(b_1 \vee b_2) &= P(b_1)+P(b_2)-P(b_1 \wedge b_2) \\ &= \tfrac{2}{3}+\tfrac{2}{3}-\tfrac{1}{3}=1 \end{aligned}$$

since $b_1 \wedge b_2$ is digit 3 or 4 and $P(b_1 \wedge b_2)=\frac{1}{3}$.

To illustrate independent and non-independent statements, take a pack of 52 playing cards (with no bias), shuffle and deal two cards. What is the chance P of two aces? If the first card is replaced before the second is dealt (after a re-shuffle), the chance of an ace at each deal is $\frac{1}{13}$. The two deals are independent, neither being affected by the other. By (iii):

$$P=\tfrac{1}{13}\,\tfrac{1}{13}=\tfrac{1}{169} \quad \text{(168 to 1 against).}$$

If the first card is not replaced, the chance of an ace at the second deal depends on whether the first card was an ace or not; the two cases giving chances of $\frac{3}{51}$ and $\frac{4}{51}$ respectively. The two deals are not independent. By (iii):

$$P=P(a_1)P(a_2 \mid a_1)$$

where a_1 is the statement: ace on first deal, and $a_2 \mid a_1$ the statement: ace on the second deal given an ace on the first. So:

$$P=\tfrac{1}{13}\,\tfrac{3}{51}=\tfrac{1}{221} \quad \text{(220 to 1 against).}$$

This is a smaller chance (longer odds against) than before.

It is to be noticed how the everyday concept of odds for or against fits the probability or chance measure as a proper fraction:

Odds: *p to q on* statement a, equivalent to chance $P(a)=\frac{p}{p+q}$

p to q against statement a, equivalent to chance $P(a)=\frac{q}{p+q}$.

A *fair bet* then corresponds to evens (1 : 1), i.e. $P(a)=\frac{1}{2}$.

Such examples seem to be intuitively obvious. The question may be raised: why develop a complicated algebra of probabilities if the results are so obvious? It is always necessary (to the mathematician) to establish strictly what may appear obvious. However, there is more to probability theory than this, even when the number of outcomes is finite. There are results which are not obvious, where intuitition can lead us seriously astray. Consider the following question.*

There have been 33 Presidents of the United States from Washington to Eisenhower inclusive. What is the chance that at least two of them have the same birthday, i.e. the same day and month? Most people would guess a rather small chance. The theory of probability shows that the chance is rather greater than $\frac{3}{4}$, i.e. 3 to 1 on. In fact, the statement that at least two Presidents have the same birthday is true; Polk and Harding were both born on 2 November.

To prove the result more generally, consider a group of r people. Assume that, for each, the 365 days in the year are equally likely as birthdays and ignore leap years. The first person has a particular birthday; the chance that the second person has a *different* birthday is $\frac{364}{365}$, the chance that the third has still a different birthday is $\frac{363}{365}$, and so on. The chance of r different birthdays:

$$\frac{364}{365}\frac{363}{365}\cdots\frac{365-r+1}{365}=\frac{364\times 363\times\ \dots\ (365-r+1)}{365^{r-1}}$$

is an application of (iii) in the non-independent case. If this proposition is called a, the proposition that at least two of the r have the same birthday is $(\sim a)$. Then $P(\sim a)=1-P(a)$ gives:

$$P_r = 1 - \frac{364\times 363\times\dots\times(365-r+1)}{365^{r-1}}$$

* Posed and analysed in Kemeny, Snell and Thompson: *Introduction to Finite Mathematics* (Prentice-Hall, 1957).

for the chance that at least two of r have the same birthday. If $r=33$, $P_r=\frac{3}{4}$ approximately. If we ask what size group makes the bet fair, we seek r for $P_r=\frac{1}{2}$. It is found that $r=23$ ($P_r=0{\cdot}51$). Hence, in betting *on* at least two people having the same birthday, the bet pays off for a group as small as 23 people. Most intuitive gamblers would be willing to bet *against* for groups larger than 23.

5.8. Finite stochastic processes. Consider a finite sequence of trials (or events, or experiments) at each of which there is a finite number of possible outcomes. If the outcomes at each stage of the sequence depend on chance, in a way to be specified in each case, then the sequence is called a finite *stochastic process.** We assume no more than the following: at a particular trial (the rth out of n) we are *given* the results of all previous trials and we then know *both* the list of possible outcomes $a_1, a_2, a_3 \ldots$ *and* their probabilities $P(a_1)$, $P(a_2)$, $P(a_3), \ldots,$ where $P(a_1)+P(a_2)+P(a_3)+\ \ldots=1$. The question to which an answer is required is: at the end of the process, what is the probability of a specified outcome, by whatever route it is reached?

A general formulation of the problem, while not difficult, involves a complicated and messy notation. Moreover, it results in little gain as compared with an analysis of each particular stochastic process as it is framed. In the following simple examples, the stochastic process in each case is represented by a *Tree Diagram,* a useful device for all classification purposes.†

(i) Six £ notes and six \$ bills are distributed between two boxes: box A has three £ notes and one \$ bill, box B three £ notes and five \$ bills. You select a box at random (by tossing a coin) and you are then permitted to select at random one note or bill from the box. What is the chance that you get a £ note? Since there are six of each, you may guess that you have an even chance, but you would be wrong. The tree diagram of Fig. 5.8a shows

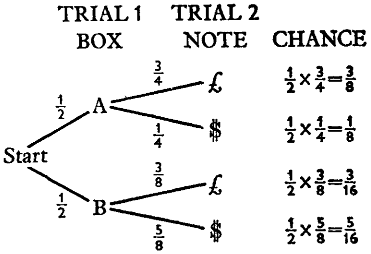


FIG. 5. 8a

* 'Stochastic' is derived from the Greek: *stochos* = guess. Roughly, it is an adjective corresponding to probability as a noun. The adjective is not 'probable' but rather 'probabilistic' if such an awkward word could be accepted.

† For other simple examples, see Kemeny, Snell and Thompson, *op. cit.*

the two trials involved. At trial 1 the outcomes are A and B, the probabilities being $\frac{1}{2}$ for each, shown on the branches leading to A and B. At trial 2, if A is the first outcome, then the outcomes are £ note and \$ bill with probabilities of $\frac{3}{4}$ and $\frac{1}{4}$ respectively, shown on the branches leading from A. If B is the first outcome, the same two outcomes are possible but with probabilities $\frac{3}{8}$ and $\frac{5}{8}$. Conditional probability, property (iii) of 5.6, then gives the chance of a final outcome. The chance of getting a £ note from box A is:

$$P(b_1 \wedge a_1) = P(a_1)P(b_1 \mid a_1) = 1/2 \times 3/4 = 3/8$$

where a_1 is 'box A selected at trial 1' and where b_1 is '£ note selected at trial 2'. The other three chances are similarly obtained. These outcomes are exclusive and their chances can be added. If a_2 is 'box B selected at trial 1' and b_2 is '\$ bill selected at trial 2', then the chances of getting a £ note or a \$ bill by either route are:

$$P(b_1) = P(b_1 \wedge a_1) + P(b_1 \wedge a_2) = 3/8 + 3/16 = 9/16$$
$$P(b_2) = P(b_2 \wedge a_1) + P(b_2 \wedge a_2) = 1/8 + 5/16 = 7/16.$$

You have a better-than-even chance of getting the £ note; it is 9 to 7 on.

(ii) Three players A, B and C are at the tennis club with darkness descending and time for only one match. Lots are drawn to decide on the odd man out, to umpire the match played by the other two. B and C are evenly matched, each having an even chance of winning. A is better, with a 75 : 25 chance of beating B or C. What chance has A of winning? The tree diagram of Fig. 5.8b shows the chances of the final outcome (by either route in each case):

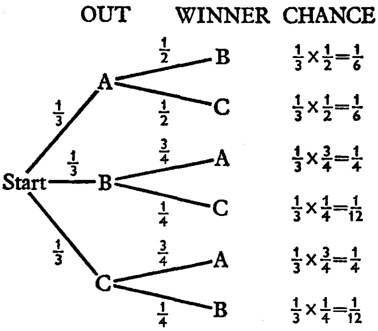


FIG. 5.8b

Winner:	A	B	C
Probability:	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$

A fair bet needs to be laid with the odds: evens on A, 3 to 1 against B and 3 to 1 against C.

(iii) You drive up to a road intersection O knowing that the place you want (W) is a short distance (but out of sight) along one of the three roads you see. The other roads lead to undesired spots A and B. You do not know which road you want and you draw lots to decide which you take (trial 1). If you decide wrongly, you return to O and draw lots again to decide between the other two roads (trial 2). If you again decide wrongly, you take the last remaining road (trial 3). The tree diagram of this stochastic process is Fig. 5.8c. In the end, the chances of making 1, 2 or 3 journeys before getting to W turn out to be equal, $\frac{1}{3}$ each.

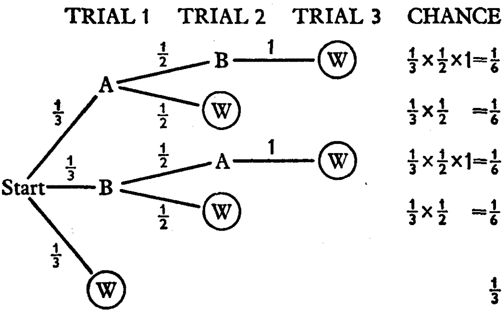


FIG. 5.8c

One more general type of finite stochastic process is of considerable importance. The process is a sequence of n trials, each being independent and having two outcomes with probabilities p and q ($p+q=1$). If $n=3$, and if one outcome is marked W (for win or success) and the other L (for loss or failure), then the tree diagram of Fig. 5.8d gives the final outcomes, according to the number of wins (out of 3):

No. wins	0	1	2	3
Chance	q^3	$3pq^2$	$3p^2q$	p^3

Similar results can be obtained for $n=4, 5, \dots$ and they generalise as follows (5.9 Ex. 26). In n trials, write $P(r)$ for the probability of exactly r wins (and $n-r$ losses), in whatever order they come. Then

$$P(r) = \binom{n}{r} p^r q^{n-r}$$

where $\binom{n}{r}=\dfrac{n!}{r!(n-r)!}(0<r<n)$ and $\binom{n}{0}=\binom{n}{n}=1$ (see 1.7 above).

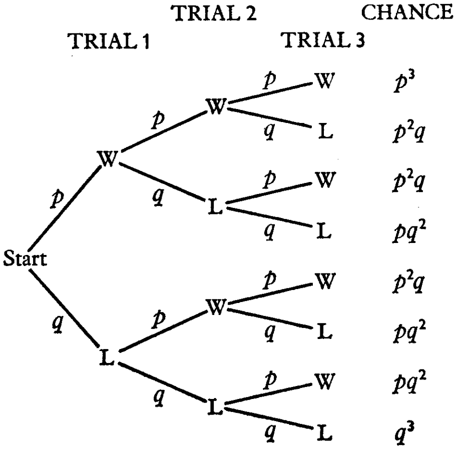


FIG. 5.8*d*

We obtain the probabilities $P(0)$, $P(1)$, $P(2)$, ... $P(n)$ for $r=0, 1, 2, \ldots n$, called the *Binomial Distribution*. The sum of the probabilities of these exclusive and exhaustive results is 1; this can be checked:

$$\begin{aligned}\sum_{r=0}^{n} P(r) &= q^n+\binom{n}{1}pq^{n-1}+\binom{n}{2}p^2q^{n-2}+\ \ldots\ +p^n\\ &=(q+p)^n \quad \text{by the } \textit{Binomial Theorem*}\\ &=1 \quad \text{since } p+q=1.\end{aligned}$$

For given n, there are $(n+1)$ values of $\binom{n}{r}$ as r ranges 0, 1, 2, ... n. Display these values in a row and write successive rows for successive values of the given n ($n=1, 2, 3, \ldots$). The values of $\binom{n}{r}$ then appear very conveniently in a triangular form, named after Pascal (1623–1662). The construction of Pascal's triangle follows from the result (5.9 Ex. 27):

* The Binomial Theorem of elementary algebra is usually written:

$$(1+x)^n=1+\binom{n}{1}x+\binom{n}{2}x^2+\ldots+x^n.$$

Here put $x=p/q$ and multiply through by q^n.

Pascal's Triangle for $\binom{n}{r}$

	$r=0$	1	2	3	4	5	6 ...
$n=1$	1	1					
2	1	2	1				
3	1	3	3	1			
4	1	4	6	4	1		
5	1	5	10	10	5	1	
6	1	6	15	20	15	6	1
...	....						

Each entry is the sum of two entries in the row above, i.e. one immediately above and one above and to the left.

Pascal's triangle provides the quickest way of finding $\binom{n}{r}$ for various n and r, when n and r are small positive integers. In the example of the Binomial Distribution ($n=3$), the coefficients of the various powers of p and q are $\binom{3}{r}$ for $r=0, 1, 2, 3$. These are read off Pascal's triangle as 1, 3, 3, 1.

5.9. Exercises

1. For p: equity prices are high and q: all equity prices are rising, translate $p \wedge \sim q$, $\sim p \wedge \sim q$, $\sim(p \vee q)$ and $\sim(\sim p \vee \sim q)$ into words.

2. Consider the statements about adults: p: the person is a man, q: the person is currently married and r: the person has been married at some time. Express the following by use of $\sim$, $\vee$, $\wedge$ and $\rightarrow$: (*a*) she is a woman, (*b*) the person is either married now or has been at some time, (*c*) the person is widowed or divorced, (*d*) the woman is widowed or divorced, (*e*) if he is married now, then he has been married at some time.

3. In Ex. 2, show that $\sim p \wedge \sim q$ refers to a single, widowed or divorced woman but $\sim p \wedge \sim r$ to a single woman.

4. In example (ii) of 5.1, show that there are three logical possibilities for rising prices as in the Truth Table:

Equity prices	q	r	$\sim q$	$\sim q \wedge r$	$\sim q \vee r$
all rising	T	T			
some, not all, rising	F	T			
none rising	F	F			

Complete the table. Deduce that $\sim q \vee r$ is logically true, but not $\sim q \wedge r$. Express in words.

5. In Ex. 4, show that $q \vee (\sim q \wedge r)$ is equivalent to r.

6. Make up a Truth Table for the six logical possibilities (man/woman; single/married/widowed or divorced) of Ex. 2 and show in it the truth or otherwise of $\sim p \wedge q$, $q \vee r$ and $q \rightarrow r$. Show that q implies r and that $q \vee r$ is equivalent to r.

*7. *The Notation* $\underline{\vee}$. Distinguish between 'p or q or both' and 'p or q but not both'. Show that, while the first is $p \vee q$, the second needs to be shown $\sim(p \wedge q) \wedge (p \vee q)$, sometimes written $p \underline{\vee} q$. Illustrate the difference with reference to p and q of example (i) of 5.1. For example (iii) of 5.1 show that $p \underline{\vee} q$ is equivalent to $\sim p \wedge q$. Interpret in words and in terms of truth sets.

8. If p and q are any statements, show that $p \vee \sim p$ is logically true, and equally $p \vee (\sim p \vee q)$. In terms of truth sets, show $P \cup P' = U$ and that

$$P \cup (P' \cup Q) = (P \cup P') \cup Q = U \cup Q = U.$$

In what sense is q here a red herring?

9. The truth set of $p \rightarrow q$ is $P' \cup Q$. Deduce that the truth set of $p \leftrightarrow q$ is $(P' \cup Q) \cap (P \cup Q') = (P \cap Q) \cup (P' \cap Q')$ and illustrate on a Venn Diagram. Show that, if the truth set of $p \leftrightarrow q$ is U, then $P = Q$ (p and q equivalent).

10. Draw Venn Diagrams to show that $p \rightarrow q$, $\sim p \vee q$ and $\sim q \rightarrow \sim p$ have the same truth set. Interpret.

11. Given: the product of an even digit with any digit is even. Prove by *reductio ad absurdum* that m and n are both odd if mn is odd.

12. Show that $(p \wedge \sim q) \rightarrow \sim p$ is equivalent to $p \rightarrow q$. Devise a method of proof of $p \rightarrow q$ which starts by taking p in conjunction with $\sim q$.

13. Show that the assertion is valid: if $p \vee q$ and $\sim q$, then p. Illustrate by truth sets.

14. Consider the argument: everything medicinal is vile, therefore everything vile is medicinal. If the first is true, show that a *necessary* condition for something to be medicinal is that it is vile; and that a *sufficient* condition for something to be vile is that it is medicinal.

15. Summarise Euclid on congruence by stating three equivalent conditions for two congruent triangles (all necessary and sufficient): three sides equal; two sides and included angle equal; one side and two angles equal. Show that the condition that two angles are equal is necessary and sufficient for similar triangles, but necessary only (not sufficient) for congruent triangles.

*16. *Reflexive sets.* Use the results of 4.7 and 4.8 to show that a necessary condition for a countably infinite set is that it is reflexive but that the Axiom of Choice is needed to establish that a necessary and sufficient condition for an infinite set is that it is reflexive.

17. Feeding pennies into a slot machine has the possible outcomes: (a) nothing back, (b) penny returned, (c) a prize won, (d) penny returned and prize won. Attempt to assign weights to the truth sets A, B, C and D given that the probability of a is $\frac{1}{2}$ and that of b or c (but not both) $5/12$. Show that you can put the odds as evens on getting something and 11–1 against getting a prize

and money back, but that you cannot write the chance of penny returned or that of winning a prize. What additional data are needed for this?

18. It is given that $P(a_1 \wedge a_2) = P(\sim a_1) = \frac{1}{4}$; $P(a_2) = \frac{1}{2}$. What is $P(a_1 \vee a_2)$?

19. In a race (won by one horse only), a bookie quotes 2 : 1 against horse A and 5 : 1 against horse B. What odds should he offer on A or B winning?

20. *Consistent statements.* Statements a_1 and a_2 are said to be consistent if $P(a_1 \wedge a_2) \neq 0$. From property (ii) of 5.6, establish that, if $P(a_1) + P(a_2) > 1$, then a_1 and a_2 are consistent.

21. There are six empty seats in a row at a cinema; three people take seats at random. What is the chance that they leave no empty seats between them? What is the chance that they leave three adjacent seats empty?

22. In example (i) of 5.4, the chance that 11 or 12 is the sum of the two digits shown by the dice is $3/36 = 1/12$. Show that it is $3/11$ given that at least one of the digits is 6, and that, if the *first* throw gives 6, it is improved to $1/3$. Why?

23. A and B are playing poker. A bets and, given his hand (a strong one), the chance of B's hand being better is $1/10$. If B has a better hand, the chance that he raises the bet is $9/10$; if B's hand is worse, the chance that he raises is $1/5$. B does raise the bet; use Bayes Theorem of 5.6 to show that the probability that B has a better hand is now to be put at $1/3$.

24. Four people leave umbrellas on the hat-stand of a restaurant. Absent-mindedly, the first three to leave pick umbrellas at random and the fourth takes the last one. Show that the chance that no one gets his own umbrella is $3/8$.

25. The audience at a small political meeting is 10 on one side of the aisle (7 Conservative, 3 Labour) and 10 on the other side (5 Conservative, 5 Labour). In taking a straw vote, the candidate tosses a coin to decide which side of the aisle to go, and then he selects at random two people, one after the other. Draw a Tree Diagram to show the outcomes and their chances. Show that the chance of getting two Conservatives is $31/90$.

26. *Binomial Distribution.* For 4 trials, draw a Tree Diagram similar to Fig. 5.8*d* and show $P(r) = \binom{4}{r} p^r q^{4-r} (r = 0, 1, 2, 3, 4)$. By induction, prove that $P(r) = \binom{n}{r} p^r q^{n-r}$ for n trials.

27. *Pascal's Triangle.* Establish the construction by showing, from the definition of $\binom{n}{r}$, that $\binom{n+1}{r} = \binom{n}{r} + \binom{n}{r-1}$ for any positive integral n and $r \leqslant n$.

28. Show that the ratio of $\binom{n}{r}$ to $\binom{n}{r-1}$ is $\frac{n-r+1}{r}$ for any positive integral n and $r \leqslant n$. Deduce that $\binom{n}{r} > \binom{n}{r-1}$ for $r = 1, 2, \dots \frac{n}{2}$ (n even). What corresponds when n is odd? Interpret in terms of a row of Pascal's Triangle.

29. In the Binomial Theorem for $(1+x)^n$, put $x = 1$ to show that $\sum_{r=0}^{n} \binom{n}{r} = 2^n$ for n a positive integer. Check from the rows of Pascal's Triangle for $n = 1$ to 6.

CHAPTER 6

GROUPS AND FIELDS

6.1. The structure of a set. In Chapters 2 and 3 we tried to disclose the basic concepts of number systems, and of sets of polynomials, and to lead towards a precise formulation of them. In such a mathematical development, there must be certain assumptions in the form of undefined properties and of definitions of terms and concepts. There are *postulates* and *definitions* to be framed. It is essential that these should be *consistent*, that nothing is in conflict with anything else. But it is also desirable that they should be elegantly and economically laid out, that they should be as concise, as simple and as general as they can be made. In particular, they should be *independent* in the sense that nothing is a consequence of anything else, and *minimal* in that they are reduced to the barest essentials.

For example, it is quite consistent to define an ordered field as something which satisfies all the rules and properties of 2.2. But it would be wasteful and not very enlightening to do so. No attention is paid to which of the rules and properties are derivable from others, nor to the minimal definitions of sets which fall short of being ordered fields. It is time to be more systematic, to simplify and to generalise concepts.

Mathematics has to do largely with the structure of a set. How are the elements related? Can they be added and multiplied? Is there an order among them? If so what are the properties involved? The properties of structure to pursue first, and mainly, are those concerned with *binary operations*, like $+$ and $\times$, by means of which one element is obtained as a combination of others. For sets of numbers, $+$ and $\times$ are ordinary addition and multiplication. For polynomials, the operations are a little more involved. In saying that $a_1x^2+b_1x+c_1$ plus $a_2x^2+b_2x+c_2$ is a third polynomial $(a_1+a_2)x^2+(b_1+b_2)x+(c_1+c_2)$, we mean that the sum rule is: the sets of three coefficients (a_1, b_1, c_1) and (a_2, b_2, c_2) add to the set $(a_1+a_2, b_1+b_2, c_1+c_2)$. Each of the three coefficients is subject to separate addition. Another and less straightforward rule applies to products of polynomials. For subsets of a given universal set (4.2) the operations are less familiar; in combining subsets we have union $\cup$ and intersection $\cap$. These have some similarity with sums and products, by no means exact but enough to make it worth while to use the notation $+$ for $\cup$ and $\times$ for $\cap$.

Binary operations, then, are of various kinds. Addition and multiplication are the most common, either in their ordinary forms or as suitable descriptions of operations which are sufficiently similar. We must keep an open mind on what operational rules will be obeyed; the rules of Boolean algebra (4.3) are no less valid or reasonable than those of ordinary algebra (2.2). From this point of view, an integral domain (like the integers) or a field (like the rationals) is a very specialised kind of set, satisfying a whole list of particular, though familiar, rules. Other sets may not be so obedient. In examining the structure of sets generally, we find it profitable to go back to something simpler, and in particular to consider first a single binary operation and not the conjunction of two such operations. This single operation may be addition or multiplication, or something more or less like one of them. We start from a completely neutral position: we take a binary operation denoted $*$, whatever it may be.

6.2. Groups. A set $\{a, b, c, \ldots\}$ is considered with respect to a binary operation $*$. The operation is a *rule of combination* applied to any two elements a and b to give another element written $a * b$. In a particular case, it must be specified precisely, e.g. by writing a table showing all combinations, as with the ordinary multiplication table. Most sets in mathematics are of a kind called 'groups', with a structure of a particularly simple but general nature in terms of the specified binary operation. This structure is that the set obeys *one* of the columns of the operational rules of 2.2, i.e. the rules in their application to one binary operation. More precisely, there are four rules to be obeyed. The first is *closure*: every pair of elements of the set must be capable of combination by the operation $*$ and the result in each case is also an element of the set. There are no exceptions and nothing new is produced. Another is the *associative* property than the order of combining three elements is immaterial: $a * (b * c) = (a * b) * c$, either being written $a * b * c$. A third is that there is an *identity* element e in the set, such that any element a is unchanged when combined with the identity element: $a * e = e * a = a$. The last is that every element has an *inverse* in the set, that each a has an inverse a^{-1}: $a * a^{-1} = a^{-1} * a = e$. When an element and its inverse are combined, the identity element results. From these properties, another follows: *cancellation*. If $a * b = a * c$, then $b = c$. Notice that nothing has yet been said about the *commutative* property that the order of combination of two elements does not matter: $a * b = b * a$. This is not laid down as a requirement for a group. It can be regarded as an extra, desirable but not essential. Many groups do obey the commutative rule; they are 'commutative groups'. There are other groups which are not so obedient.

We have not yet arrived at a strict and economical definition of a group. While the properties mentioned are all consistent, some of them can be deduced from others. This is true of cancellation as already indicated. It is also true of part of the rule for the identity; if $e * a = a$, then $a * e = a$. Similarly for the inverse: if $a^{-1} * a = e$, then $a * a^{-1} = e$. Finally, it is not necessary to lay down either that the identity e is unique or that the element a has a unique inverse a^{-1}. It is certainly desirable that there should be one and only one identity (inverse) but it is not necessary to specify it. The uniqueness follows as a consequence of the other rules. (For proofs see 15.3.)

Hence, as a strictly axiomatic definition:

DEFINITION: *A set $G = \{a, b, c, \ldots\}$ of elements of any specified kind, in which a binary operation $*$ is specified, is a* **group** *if the four postulates shown below are satisfied.*

For any elements $a, b, c, \ldots$ of a group G:

Property	*Postulates*	*Operational Rules*
Closure	$a * b \in G$	$a * b \in G$
Associative	$a * (b * c) = (a * b) * c$	$a * (b * c) = (a * b) * c = a * b * c$
Commutative	......	$a * b = b * a$ may be true for all $a, b \in G$ (commutative group) or it may not (non-commutative group)
Identity	there is an element $e \in G$ such that $e * a = a$	the identity e is unique and $a * e = e * a = a$
Inverse	there is an element $a^{-1} \in G$ such that $a^{-1} * a = e$	the inverse a^{-1} is unique and $a * a^{-1} = a^{-1} * a = e$
Cancellation	......	if $a * b = a * c$, then $b = c$

The operational rules shown comprise the four postulated properties with the addition of others which follow from the postulates. They are in line with the rules of 2.2.

From the properties of number systems, it is seen that the integers, rationals, real and complex numbers are all groups under addition. The set of positive integers is not a group since there is no identity (zero) and no inverses (negatives). If the operation considered is multiplication, then the rationals, real and complex numbers are all groups. However, the set of integers is not a group under multiplication; there is an identity (the integer 1) but no other integer has an inverse (reciprocal). Similarly, the set $F[x]$ of polynomials is a group under addition, but not a group under multiplication since again reciprocals are lacking.

A group $G=\{a, b, c, \dots\}$ has all kinds of subsets. The question is: does a subset of G itself satisfy the four postulates for a group. If so, it is a group in its own right, and called a *subgroup*. It is necessary that the elements of a subgroup K of a group G satisfy all the four postulates of a group. On the other hand, we can get by with less since many of the properties of G carry over into the subset. It is sufficient to write two conditions only:

THEOREM: *A subset K of a group $G=\{a, b, c, \dots\}$ is a* **subgroup** *under the conditions: (a) if $a, b \in K$, then $a * b \in K$; (b) if $a \in K$, then $a^{-1} \in K$. A subgroup K of G must contain the identity e of G.*

Proof: if K is a subgroup then the four postulated properties hold and these include (a) and (b). On the other hand, if (a) and (b) hold, then K is closed under $*$ by (a) and two applications give: if $a, b, c \in K$, so do $a * (b * c)$ and $(a * b) * c$. These are the same since they are so in G itself (associative property for K). Again, if $a \in K$ then (b) gives $a^{-1} \in K$ and (a) gives $a^{-1} * a \in K$. But $a^{-1} * a = e$ in G and so $e \in K$. This, with the fact that $a^{-1} \in K$, shows that the identity and inverse properties hold for K. Hence, by this process of checking each postulate (property) in turn, it follows that K is a group. Q.E.D.

It is to be noticed that a set of a single element $\{e\}$ is always a group, for e serves as the identity and its own inverse: $e * e = e$. Further, a group G has at least one subgroup, i.e. the set of one element $\{e\}$.

The following three examples of *groups under addition* serve to indicate the variety of sets which are groups. There are infinite groups, as in example (i), and finite groups as in (ii). There are groups of numbers and groups of other kinds of elements as in (iii). In all practical cases, groups under addition are commutative and they can then be called simply *additive groups*. When the operation is addition, the symbol * used above is replaced by $+$ and the identity e by zero (0). The inverses in the group are then the negatives and written $a^{-1} = -a$ where $(-a)+a=0$. Since negatives exist, the inverse operation of *subtraction* can be performed: $a-b=a+(-b)$.

(i) *The set J of integers.* This is a commutative group under addition since all the rules (including the commutative one) are obeyed. The identity is zero 0 and the inverse of any integer m is the negative $(-m)$. Of the subsets of J, many are not groups, including the subset J^+ of positive integers which lacks an identity (zero) and inverses. The subset of even integers $\{\ldots\ -4,\ -2,\ 0,\ 2,\ 4,\ \ldots\}$ is a group, a subgroup of J. This follows since the conditions (a) and (b) of the Theorem above are satisfied; the sum of two even integers and the negative of an even integer are even integers. The same is true of the subset of J consisting of all multiples of 3, or indeed of any integer. See 6.9 Exs. 2 and 12.

(ii) *The set of integers* (*mod n*). This is a commutative group under addition. The sum of two or more integers is a member of the set and the order of adding is immaterial (associative and commutative). The identity is 0, the inverse of 0 is 0, and the inverse of m ($\neq 0$) is $n-m$, since $m+(n-m)=n=0$ (mod n). Consider $\{0,\ 1,\ 2,\ 3,\ 4\}$ (mod 5) as an instance. In a finite set with a specified (and relatively small) number of elements, the table of the operation can be written and used directly to demonstrate whether or not the set is a group. Here:

+	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

That 0 is the identity is seen from the (unchanged) elements in the row 0. That each element has an inverse (negative) is seen from the fact that 0 appears just once in each row, the corresponding elements being each other's negatives (e.g. 1 and 4; 2 and 3). That the group is commutative is seen from the fact that the table is symmetric about the leading diagonal.

(iii) *The set S of all subsets* of a given universal set U. If the union of two subsets is defined (as in 4.2) as all elements in either or both of the subsets, then it is seen at once that S is not a group under addition (union). From the operational rules for sets, if $S=\{A, B, C, \ldots\}$, then:

$$A+B \in S \text{ (closure)}; \quad A+(B+C)=(A+B)+C \text{ (associative)};$$
$$A+0=A \text{ (identity)}$$

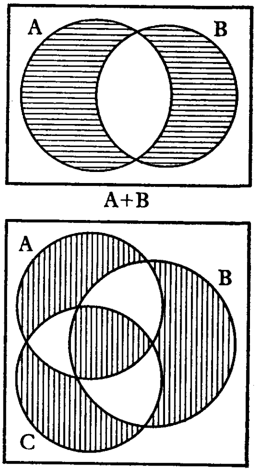
all hold for union (+), the empty set being written 0 (zero). But there is no inverse of any given (non-empty) subset A; since $A+B=0$ is not possible, there is no subset B to serve as the negative of A.

However, addition (union) can be *re-defined* so that the properties of closure, associative and identity still hold and so that the inverse property also holds. The addition (union) of A and B is taken as the set $(A+B)$ of all elements in A or in B but *not* in both, as shown in the Venn Diagram of Fig. 6.2. The figure also shows that the associative property is still true, as also are closure and the identity (the empty set 0). Further:

$$A+A=0.$$

Each subset A is its own inverse: $(-A)=A$.

The set $S=\{A, B, C, \ldots\}$ is a group under addition. The peculiar feature is that each subset A serves as its own negative. Hence, subtraction and addition are by definition the same: $A-B=A+B$.



A+B

A+(B+C)=(A+B)+C

FIG. 6.2

Of even more interest are *groups under multiplication*. When the operation is the product of two elements, the neutral symbol $*$ is replaced by × (which can then be dropped when there is no ambiguity) and the identity e becomes unity (1). Sets of numbers may or may not be groups under ×, as the examples below show, but if they are then the commutative property also holds. Groups of numbers are commutative groups. The same is not necessarily true of groups of other kinds of elements under ×, which is why groups are so interesting.* The illustration of non-commutative groups is so important that it is left for separate discussion in the following two sections. Any group under × (commutative or not) is such that every element a has an inverse or reciprocal a^{-1}. But it is only for a commutative group that the inverse operation of *division* can be uniquely performed: $\frac{a}{b}=ab^{-1}=b^{-1}a$.

(iv) *The set R of rationals* (excluding 0). This is a commutative group under multiplication since all the rules, including the commutative one, are obeyed. Of subsets of R, the set J of all integers is an example of one which is not a multiplicative group since it lacks reciprocals; J is a group under addition but *not* under multiplication. On the other hand, the set $\{2^n \mid n \text{ an integer}\}$ is a group under multiplication, but *not* under addition.

(v) *The set* $\{-1, 1\}$. This is the simplest of all groups under ×, apart from the case of a set comprising the element 1 alone. It is a group since all the four rules hold, each element being its own reciprocal; it is also commutative. It is another subgroup of the group R. It is also a subset of J, i.e. J itself is not a group under × but does have a subset which is a group.

(vi) *The set of integers (mod n)*, again excluding 0. This is a group under × if n is prime but it is not a group otherwise. The general result (6.9 Ex. 9) can be illustrated in two cases where the multiplication tables, shown here, demonstrate whether or not the group properties hold. For $n=4$, they do not; there is a failure to jump the first hurdle. The set is not closed since $2\times 2=0$; not an element of the set. The set also lacks reciprocals since there is no r for $2\times r=1$, i.e. 2 has no reciprocal. The set of integers (mod 4) is not a group under × even when 0 is excluded. For $n=5$, the set is far more obedient. It is

$n=4$

×	1	2	3
1	1	2	3
2	2	0	2
3	3	2	1

$n=5$

×	1	2	3	4
1	1	2	3	4
2	2	4	1	3
3	3	1	4	2
4	4	3	2	1

* Note that, if $G=\{a, b, c, \ldots\}$ is a non-commutative group under ×, then $a\times b=b\times a$ is not generally true, i.e. not true of *all* a and b in G. It is true of *some* a and b and, in particular, it is true of the identity ($a\times 1=1\times a$) and of reciprocals ($a\times a^{-1}=a^{-1}\times a$).

closed, satisfies the associative and commutative rules, has an identity 1 and any member has a reciprocal. The reciprocal is to be found by seeking an entry 1 in any row of the multiplication table; and every row has just one entry 1. The integers (mod 5) with 0 excluded form a commutative group under multiplication. These results are a re-interpretation of those of 2.7 above.

(vii) *The set S of all subsets* of a universal set. Consider the operation of products (intersection) with the identity element 1 taken as the universal set. The operational rules for sets show that the properties of a group hold, except that reciprocals are lacking. If A is a proper subset of the universal set, there is no subset B such that $A \times B = 1$. S is not a group under multiplication.

The group $\{-1, 1\}$ corresponds to handling odd and even, writing odd $= -1$, even $= 1$ and taking products. A combination of odd and even is odd $(-1 \times 1 = -1)$; other combinations are even $(-1 \times -1 = 1 \times 1 = 1)$. It is also the simplest case of a 'cyclic' group:

DEFINITION: *A group G under multiplication is* **cyclic** *if the elements $\{a, a^2, \dots a^{n-1}, a^n\}$ are generated by powers of a single element a, which is such that $a^n = 1$ (identity), for some positive integer n.*

Note that the finite set of elements is closed, as it must be for a group:

$$a^s a^t = a^{s+t} = a^{qn+r} = (a^n)^q a^r = 1^q a^r = a^r \in G$$

where s and t are positive integers $(\leqslant n)$, where $s + t = qn + r$ having remainder r $(\leqslant n)$ and where $a^n = 1$ is the essential property to ensure closure. All the other properties of a group are satisfied, including the commutative rule.

Cyclic groups are not as uncommon as might be thought. Consider the following cyclic groups of elements from the field of complex numbers. Put $a = -1$, giving the cyclic group $\{-1, 1\}$. Put

$$a = \omega = \tfrac{1}{2}(-1 + i\surd 3),$$

so that $\omega^2 = \frac{1}{2}(-1 - i\surd 3)$ and $\omega^3 = 1$, giving the cyclic group $\{\omega, \omega^2, \omega^3\}$. Put $a = i$, giving the cyclic group $\{i, -1, -i, 1\}$. The essential property in the last case is $i^4 = 1$, from the definition $i^2 = -1$. These examples are the cyclic groups of the nth roots of unity, for $n = 2, 3, 4$. They suggest that the nth roots of unity are a cyclic group for any positive integral n. This is so, from the theorem of 3.8. There are also cyclic groups under addition (6.9 Ex. 8).

6.3. Transformations. The object here is to illustrate some of the great variety of types of transformations used in diverse branches of mathematics and to view them from one aspect: a set of transformations as a group.* The group operation is multiplication, taken as two transformations applied one after the other in succession.

(i) *Translations and magnifications of a figure.* Start with a given figure, say a square located in a plane (A in Fig. 6.3). Stretch the figure in a given direction (horizontally in B of Fig. 6.3) and let the stretch be in a given ratio $a : 1$ ($a > 0$). This is transformation of the square into a rectangle, a magnification in one direction. There are various transformations according to the values of a ($a > 1$, magnification; $a < 1$, squeeze; $a = 1$, no change). Alternatively, take the square and shift it bodily in a given direction (horizontally in C of Fig. 6.3) and let the shift be a given distance b. This is a transformation of the square into another square, a translation in one direction. The value of b can vary ($b > 0$, shift to right; $b < 0$, shift to left; $b = 0$, no change).

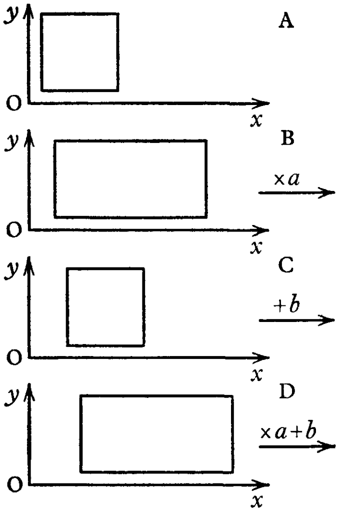


FIG. 6.3

Given two transformations, they are combined by applying one and then the other. Two magnifications, first by $a : 1$ and then by $\alpha : 1$, give a combined magnification of $a\alpha : 1$. Here the order does not matter (commutative). Two shifts, first by b and then by β, give a combined shift of $b + \beta$, and again the order does not matter. The position is different when a magnification and a shift are combined and it is most easily explored in algebraic terms.

Fix axes Ox and Oy in the plane and units for measurement along Ox and Oy. Any point (e.g. the corner of the given figure) is then shown by a pair of co-ordinates (x, y). If the transformations are in the direction Ox, then they leave y unchanged but change x into x' according to the rules:

* Another approach to transformations is followed in 7.5 below.

(1) Magnification by $a : 1$ $\quad x' = ax$.

(2) Translation by $+b$ $\quad x' = x + b$.

Apply (1) and then (2), giving a combined transformation (2) × (1), where the order is read from right to left (see 1.7). This is from x, through x' to x'':

$$x' = ax \quad \text{and} \quad x'' = x' + b \quad \text{give} \quad x'' = ax + b.$$

This is a more general transformation, a magnification and a shift together, illustrated in D of Fig. 6.3. Now apply (1) and (2) in reverse order:

$$x' = x + b \quad \text{and} \quad x'' = ax' \quad \text{give} \quad x'' = ax + ab.$$

Hence the transformation (1) × (2) is not the same as the transformation (2) × (1); they are different examples of the general type (magnification and shift). Transformations (1) and (2) are not commutative.

The transformations are interpreted in terms of moving a figure, the axes and units being unchanged. They can be interpreted equally well in terms of a fixed figure, the axes and units being changed. The magnification by $a : 1$ is then a change in the unit for measuring x (e.g. $a = 3$, changing from yards to feet). The translation by $+b$ is a change in the origin for measuring x (e.g. distances in miles E. of London on old scale, E. of New York on new scale). The general (magnification and shift) transformation is a change in both unit and origin for x; e.g. $x' = 32 + 9x/5$ changes the measure of a temperature from $x°$ C. to $(x')°$ F.

(ii) *Movements of a figure.* The rotation of a figure can be handled on exactly the same lines as in (i), involving a transformation from co-ordinates (x, y) to co-ordinates (x', y'). (See 6.9 Ex. 15.) Consider, however, a rather different geometrical approach, when the figure to be moved is regular, e.g. an equilateral triangle, a square, or (generally) a regular polygon. The case of the equilateral triangle $\triangle$ suffices for illustration. Another case is given in 6.9 Ex. 16.

Cut $\triangle$ out of cardboard and consider all the ways in which the piece can be put back in the hole left in the cardboard. If the original $\triangle$ has vertices lettered A, B and C, there are six ways of moving $\triangle$ into place, as shown in the table:

$\begin{matrix} A \\ B\ C \end{matrix} \rightarrow$					
1:	$\begin{matrix} A \\ B\ C \end{matrix}$	ω:	$\begin{matrix} C \\ A\ B \end{matrix}$	ω^2:	$\begin{matrix} B \\ C\ A \end{matrix}$
p:	$\begin{matrix} A \\ C\ B \end{matrix}$	q:	$\begin{matrix} C \\ B\ A \end{matrix}$	r:	$\begin{matrix} B \\ A\ C \end{matrix}$

Here: 1 — leave △ unchanged

ω — rotate anti-clockwise through 120° (A to B, B to C and C to A)

ω^2 — rotate anti-clockwise through 240° (A to C, B to A and C to B)

p — turn over about the perpendicular from A to BC

q — turn over about the perpendicular from B to AC

r — turn over about the perpendicular from C to AB.

Each of these is a transformation of △ into another position in which the vertices are different. The set of six can be called the *movements* of the triangle.

The first three (1, ω and ω^2) correspond to rotations and they combine easily among themselves. In particular:

$$\omega \text{ followed by } \omega: \begin{matrix} A \\ B\ C \end{matrix} \rightarrow \begin{matrix} C \\ A\ B \end{matrix} \rightarrow \begin{matrix} B \\ C\ A \end{matrix} = \omega^2 .$$

Hence $\omega \times \omega$ (two successive transformations) is ω^2, i.e. two rotations of 120° gives one rotation of 240°. This is the justification for the use of the notation ω^2. Further:

$$\omega \text{ followed by } \omega^2: \begin{matrix} A \\ B\ C \end{matrix} \rightarrow \begin{matrix} C \\ A\ B \end{matrix} \rightarrow \begin{matrix} A \\ B\ C \end{matrix} = 1$$

$$\omega^2 \text{ followed by } \omega: \begin{matrix} A \\ B\ C \end{matrix} \rightarrow \begin{matrix} B \\ C\ A \end{matrix} \rightarrow \begin{matrix} A \\ B\ C \end{matrix} = 1$$

i.e. this double application is commutative and $\omega \times \omega^2 = 1$. A rotation of 120° and one of 240° combine to give a rotation of 360°, which leaves △ unchanged. The three transformations 1, ω, ω^2 are commutative and, clearly, linked with the cube roots of unity in some way.

The other three (p, q, r) can be combined with themselves or with the first three, and the results are non-commutative. For example:

$$\omega \text{ followed by } p\text{: } \begin{array}{c}A\\ B\ C\end{array} \to \begin{array}{c}C\\ A\ B\end{array} \to \begin{array}{c}C\\ B\ A\end{array} = q$$

$$p \text{ followed by } \omega\text{: } \begin{array}{c}A\\ B\ C\end{array} \to \begin{array}{c}A\\ C\ B\end{array} \to \begin{array}{c}B\\ A\ C\end{array} = r.$$

The complete set of pairs of transformations can be worked out and tabled:

×		*First transformation:*					
		1	ω	ω^2	p	q	r
Second	1	1	ω	ω^2	p	q	r
transf.:	ω	ω	ω^2	1	r	p	q
	ω^2	ω^2	1	ω	q	r	p
	p	p	q	r	1	ω	ω^2
	q	q	r	p	ω^2	1	ω
	r	r	p	q	ω	ω^2	1

(iii) *Permutations of a collection of objects.* Consider the case of three objects, placed in an original order: (A, B, C). There are six permutations:

$$(A, B, C)\to p_1\text{: } (A, B, C) \quad p_2\text{: } (C, A, B) \quad p_3\text{: } (B, C, A)$$
$$p_4\text{: } (A, C, B) \quad p_5\text{: } (C, B, A) \quad p_6\text{: } (B, A, C).$$

Each of these is a transformation of the collection of objects from one order to another. They can be combined, e.g.

p_2 followed by p_4: $(A, B, C)\to(C, A, B)\to(C, B, A)=p_5$
p_4 followed by p_2: $(A, B, C)\to(A, C, B)\to(B, A, C)=p_6$.

Hence two successive permutations is itself a permutation, but the process need not be commutative. There is, however, no need to pursue further since this set of six transformations is essentially the same as the set of six transformations in (ii), with $p_1=1$, $p_2=\omega$, $p_3=\omega^2$, $p_4=p$, $p_5=q$, $p_6=r$. It is only a matter of interpretation: the movements of the triangle correspond to the permutations of the labels (A, B, C) attached to the vertices.

6.4. Groups of transformations. From the examples given, it is clear that transformations can be arranged in sets. There is, from (i) of 6.3, the infinite set of magnification/shift transformations $x'=ax+b$ for various real values of a and b. There is, from (ii) or (iii), the finite set of six movements of a triangle, or six permutations of three objects. The elements of such a set can be combined, one transformation being applied after another. There is, therefore, a binary operation defined in the set; it can be called multiplication and denoted ×, though it is to be interpreted as successive applications of the named transformations. The product of any two transformations may or may not be commutative. The question naturally arises: is the set of transformations a group?

The range and depth of the concept of a set and group are greatly increased. We have met sets and groups of numbers, or of number sequences (e.g. polynomials). We now have sets and groups of operations or transformations. The elements which compose the sets are now something quite different. Wide new vistas are opened up. Historically, indeed, the concept of a group was developed first for transformations, and later generalised to other entities such as numbers or polynomials. We have here simply reversed the historical order.

A definition and a notation are required for transformations of the kinds considered (and many others):

DEFINITION: *A* **transformation** *is the result of applying a specified* **operator** *r to certain objects A, B, C, ...*

NOTATION: *A transformation T is written $A \xrightarrow[T]{} r(A)$, where the operator r applied to A gives $r(A)$.*

For example, A may be a triangle and r may be the operator 'rotate through 120°'; or A may be a collection of three things in order and r may be the operator 'permute the collection by interchanging the second and third things'. The terms 'operator' and 'transformation' are not always kept distinct. It is as well to reserve the term 'operator' for the rule whereby one object is changed or transformed into another object, and the term 'transformation' to the whole process of selecting A and of getting the transformed object $r(A)$.

The product of two transformations (operators r and s) is defined:

DEFINITION: *The product of two transformations is the result of their successive application. If r is first applied to A and then s applied to the result $r(A)$, the product is: $s(r(A)) = sr(A)$.*

Here $sr(A)$ is to be interpreted: r first, then s. The other product $r(s(A))$, or more simply $rs(A)$, is to be interpreted: s first, then r.

The operators are put in front of the object and read from right to left.*

Consider a set of transformations: $r(A), s(A), t(A), \dots$ in which products are defined as the double transformations such as $sr(A)$ or $rs(A)$. The products need not be commutative: $sr(A) \neq rs(A)$. A different transformation may arise from r first, s second, than from s first, r second. To establish that we have a *group of transformations*, we must check that the four postulates for a group are satisfied. The set must be closed: $sr(A)$ must itself belong to the set if $r(A)$ and $s(A)$ belong. Products must be associative: $t(sr)(A) = (ts)r(A)$, i.e. the double operator sr first and t second must give the same result as r first and the double operator (ts) second. There must be an identity 1, i.e. an operator 1 which leaves the object unchanged and which therefore gives $r \times 1(A) = 1 \times r(A) = r(A)$. Finally, there must be inverse (reciprocal) transformations: each operator r has an inverse r^{-1} such that $rr^{-1}(A) = r^{-1}r(A) = A$. The inverse transformation undoes what is done by the original transformation. Two examples illustrate:

(i) *The group of translations* of a figure. The translation $b(A)$ shifts the figure A a distance b to the right. If any point in the plane (e.g. a corner of the figure A) has co-ordinates (x, y) referred to axes Oxy (Ox horizontal), then the transformed point has co-ordinates (x', y) where $x' = x + b$. This is a single transformation (which may be applied to various figures $A, B, C, \dots$) if b is a specified value. A set of transformations is obtained as b varies. Suppose $b = n$, a positive, zero or negative integer. The set of translations is then $n(A)$, i.e. $x' = x + n$, as n ranges over the set J of integers. The product of two translations is $mn(A) = nm(A)$, where successive shifts of n and of m (or conversely) result in a shift of $(m+n)$. The first shift gives $x' = x + n$, the second gives $x'' = x' + m$, and the product of the two shifts gives $x'' = x + (m+n)$. Hence, the properties of the set of translations $n(A)$ under products are paralleled by the properties of the set J of integers under sums. J is a group under addition and so is the set of translations $n(A)$ under multiplication. Both are commutative.

(ii) *The group of the equilateral triangle.* The set is a finite one,

* There is an alternative notation, sometimes used and giving rise to some confusion, in which the operators are put behind the object and read from left to right: $A(rs)$ for $(Ar)s$, or first r operating on A, then s on the result.

consisting of six transformations (operators 1, ω, ω^2, p, q, r) applied to the triangle $\triangle$ as in example (ii) of 6.3. The set is found to be a non-commutative group. The result applies equally to the set of six permutations of a collection of three things, as given in example (iii) of 6.3; this set of permutations is the same non-commutative group.

Products of transformations (operators in succession) are defined according to the table of 6.3. So: $p\omega(\triangle)=q(\triangle)$ and $\omega p(\triangle)=r(\triangle)$, illustrating that products are not commutative. They are closed and associative; for example:

$$\omega^2(p\omega)(\triangle)=\omega^2 q(\triangle)=r(\triangle)$$
$$(\omega^2 p)\omega(\triangle)=q\omega(\triangle)=r(\triangle).$$

There is an identity, the operator 1 leaving $\triangle$ unchanged. Finally, each transformation has its inverse, given by the operator which undoes what the original one does. Whenever an entry 1 appears in the table, the two operators concerned are inverse. Since each row has just one entry 1, there are unique inverses: ω and ω^2 are inverse to each other; p, q and r are each its own inverse. Hence the set is a non-commutative group. There is also a subgroup of three transformations (operators 1, ω, ω^2), as is seen at once from the table. The product of any two of these three is also one of the three. The inverse of any one of the three is also one of the three. These are the conditions for a subgroup by the theorem of 6.2.

The subgroup (1, ω, ω^2) corresponds to rotations of $\triangle$ through 0°, 120°, 240° respectively. Unlike the complete group of six movements, the subgroup is commutative and cyclic. Take powers of the rotation ω (through 120°):

$$\omega\omega(\triangle)=\omega^2(\triangle);\quad \omega\omega\omega(\triangle)=\omega\omega^2(\triangle)=1$$

i.e. the subgroup consists of ω, its square ω^2 and its cube $\omega^3=1$. These are rotations through 120°, 240° and 360°, the last being the same as no change. The three operators can be associated with the cube roots of unity (1, ω, ω^2), where $\omega=\frac{1}{2}(-1+i\surd 3)$, complex numbers representing the vertices of an equilateral triangle on an Argand Diagram (as shown in 3.8). Multiplication by the complex number ω moves one vertex of the triangle in the diagram to the next (anti-clockwise), and this is the same as rotating the triangle through

120° (operator ω). Everything links together neatly. The cyclic group, ω, ω^2, $\omega^3=1$, can be used either for the complex cube roots of unity, or for rotations of an equilateral triangle.

6.5. Fields. We are primarily interested in sets when (and because) they have the structure of a group in respect of some binary operation. Some sets have only one such operation to consider, e.g. sets of transformations under multiplication. Other sets have two such operations, usually sums and products, e.g. sets of numbers or polynomials. These are systems of *double composition*: a set $\{a, b, c, \ldots\}$ in which a sum $(a+b)$ and a product ab is defined for each pair of elements. The questions which arise are: is the set a group from the point of view of each of the operations by itself, and how are the operations related?

Here we consider the most obedient of all systems of double composition, that described as a *field*. As shown in 15.3, the concept of a field can be derived from the combination of two groups (one under $+$ and the other under $\times$) with a distributive property linking $+$ and $\times$. The development proceeds by first defining the more general (less specialised) concept of a 'ring' and then by adding extra properties until the less general (more specialised) field is obtained. Alternatively, an independent definition can be written for a field, specifying the properties of sums and products (closure, associative, commutative), the distributive rule connecting them, and the requirement that equations of the simple kind: $x+a=b$ and $ax=b$ should have solutions. The group properties of a field are then deduced. The end result of the two procedures is the same (see 15.3). Here, having already developed the group concept, we now write quite simply:

DEFINITION: *In a set $F=\{a, b, c, \ldots\}$ of elements of any kind, two binary operations give sums $(a+b)$ and products ab. F is a* **field** *if*

(1) *the elements form a commutative group $(F+)$ under addition with identity zero* (0)*;*

(2) *the elements other than zero form a commutative group $(F\times)$ under multiplication with identity unity* (1)*;*

(3) *sums and products are distributive: $a(b+c)=ab+ac$.*

A field is a set which unites in itself two groups, the additive and the multiplicative, and in which the distributive rule links the two operations.

Several features are to be noticed. Firstly, there are two particular elements of F, zero (0) and unity (1). By use of 0, the additive group gives negatives and hence *subtraction*: $a-b=a+(-b)$. By use of 1, the multiplicative group gives reciprocals and hence *division*: $a/b=ab^{-1}$. Any element can be multiplied by 0 (i.e. $a\times 0=0$, see 6.9 Ex. 20) but 0 must itself be excluded from the multiplicative group. The reciprocal b^{-1} and quotient a/b can only be written if $b\neq 0$. Secondly, the existence of inverses (negatives and reciprocals) implies *cancellation*:

$$\text{If } a+b=a+c, \text{ then } b=c; \quad \text{if } ab=ac\ (a\neq 0), \text{ then } b=c.$$

In particular, there are no zero divisors: if $ab=0$, then $a=0$ or $b=0$. Thirdly, only one form of the distribution rule is given. Another form follows since F has the structure of a commutative group:

$$(a+b)c=c(a+b)=ca+cb=ac+bc.$$

The dual distributive property, as valid for Boolean algebra, does *not* hold for a field, i.e. $a+bc\neq(a+b)(a+c)$.

As a result, a field F satisfies all the operational rules of 2.2 with no exceptions whatever. Conversely, any set which satisfies all the rules is a field. If the set falls short of this, then some operational rule or other goes by the board.

Just as groups have subgroups, so a field has all kinds of subsets of which some can be fields in themselves (i.e. subfields). Necessary and sufficient conditions for a subset K of a field F to be a subfield stem from the conditions for a subgroup (6.2). A form of the conditions which is easy to use in practice is: (a) if $a, b\in K$, then so do $a+b$ and ab, and (b) if $a\in K$, so do $-a$ and a^{-1} ($a\neq 0$). A neater form is given in 6.9. Ex. 21. In particular, a subfield must contain both 0 and 1, the identities of F. One field can be extended into and contained within a larger field; and, in its turn, the larger field can be extended again into a still larger one. The process of *adjunction* of a new element as in 2.3, is useful in this connection. Conversely, a given field may contain a subfield, the subfield a further subfield, and so on. In such a set of 'Chinese boxes', we may look for the smallest subfield, the inner-most box. Such a subfield, called a *prime field,* has no proper subset which is itself a field.

Though highly specialised, a field is a set which quite commonly occurs in mathematics and it is the most practical and obedient kind of set we have:

(i) *The set R of rationals.* All the operational rules are valid, for ordinary sums and products, and R is a field. No subfield can be found, i.e. R is a prime field. In particular, the set J of integers is *not* a field, since it lacks reciprocals. J is the kind of 'ring' (an integral domain) which is a very near approach to a field, without quite making it. The development of 2.6 shows how the integral domain J is turned into the field R by defining fractions and making good the lack of reciprocals. A term used in the past for a field is 'domain of rationality', expressing this idea.

(ii) *The set* $R(\surd 2)=\{a+b\surd 2 \mid a \text{ and } b \text{ rationals}\}$. This is a field, obtained by adjunction of $\surd 2$ to R (2.3). It has a subfield, i.e. R. The reason why $a+b\surd 2$ provides a closed set, not only for sums and differences (which is obvious enough) but also for products and quotients, is because:

$$(a+b\surd 2)(c+d\surd 2)=(ac+2bd)+(ad+bc)\surd 2$$

and
$$\frac{a+b\surd 2}{c+d\surd 2}=\left(\frac{ac-2bd}{c^2-2d^2}\right)+\left(\frac{bc-ad}{c^2-2d^2}\right)\surd 2$$

as shown in Appendix A.6. So we always keep within the form $a+b\surd 2$.

(iii) *The sets of real and complex numbers* are both fields, satisfying all operational rules. The Chinese boxes are building up from the prime field R:

$$R\subset R(\surd 2)\subset R^*\subset C.$$

Instead of $R(\surd 2)$, we could substitute fields obtained by adjunction of other surds. C is obtained by adjunction of i to R^*:

$$C=R^*(i)=\{a+ib \mid a \text{ and } b \text{ real}\}.$$

Again closure is obtained with the form $a+ib$, for a similar reason to that just given for $a+b\surd 2$. Here:

$$(a+ib)(c+id)=(ac-bd)+i(ad+bc)$$

and
$$\frac{a+ib}{c+id}=\left(\frac{ac+bd}{c^2+d^2}\right)+i\left(\frac{bc-ad}{c^2+d^2}\right).$$

(iv) *The set* $F(x)$ *of rational fractions*

$$=\left\{\frac{f(x)}{g(x)}\mid f(x) \text{ and } g(x) \text{ polynomials}\right\}$$

is a field, obtained from the integral domain $F[x]$ of polynomials over the field F in the same way that R is obtained from J.

(v) *The set of integers (mod n).* Examples (ii) and (vi) of 6.2 give us what we need, apart from checking that + and × are connected by the distributive rule. The check is simple, since integers are so connected and since taking remainders on division by n makes no difference here (6.9 Ex. 18). Hence, if n is prime, the integers (mod n) are a field. If n is not prime, the set falls considerably short of being a field; reciprocals are lacking and there are zero divisors. For example, in {0, 1, 2, 3} (mod 4), 2 has no reciprocal and $2\times 2=0$.

+	Even	Odd
Even	Even	Odd
Odd	Odd	Even

+	0	1
0	0	1
1	1	0

×	Even	Odd
Even	Even	Even
Odd	Even	Odd

×	0	1
0	0	0
1	0	1

(vi) *The set {Even, Odd}.* The operations of + and × are defined by the tables shown in comparison with the tables for the set {0, 1} (mod 2). The latter set is a field since 2 is prime. The set {Even, Odd} is an exact parallel, and also a field. We have the simplest kind of field, of two elements, one being zero (even) and the other unity (odd). Negatives come from $1+1=0$, i.e. 1 (odd) is its own negative. Reciprocals come from $1\times 1=1$, i.e. 1 (odd) is its own reciprocal. It may be noticed, in passing, that the set $\{1, -1\}$ under × can be associated with {0, 1} (mod 2) or with {Even. Odd} under +. The association does not work for the other operation. $\{1, -1\}$ is not a field; it is not closed and has no zero under addition.

6.6. Algebraic numbers. An interesting, if rather academic, question is the following. All zeros of all polynomials with rational coefficients are complex numbers; how many different ones are there? First, since every rational number α is the zero of some polynomial (e.g. the linear polynomial $x-\alpha$), the set of all zeros includes the *countably infinite* set R of rationals. On the other hand, since every zero is a complex number, the set of all zeros is included within the non-countably infinite set C. It is easily seen that the set of all zeros is a field, closed for sums and products, and that it is countably infinite like the rationals (see 4.9 Ex. 22). This is a quite remarkable result. Numbers which are roots of polynomial equations are called *algebraic numbers*; they form a *countably infinite field* A.

The field A fits in between the prime field R and the larger field C; A is a subfield of C and R a subfield of A: $R \subset A \subset C$. We know already that the field R^* of real numbers also fits in a similar way: $R \subset R^* \subset C$. The question is: how are A and R^* related? They are both subfields of C; both of them have a prime subfield R. First, we know that the algebraic numbers of A, as roots of polynomial equations, include all rationals, at least some irrationals (like $\sqrt{2}$) and at least some non-real numbers (like i). Hence the intersection $A \cap R^*$ is not empty and it is a proper subset of A: $(A \cap R^*) \subset A$. $A \cap R^*$ comprises the *real algebraic numbers*, the real roots of polynomial equations. The rest of A is made up of the non-real (complex) roots of polynomials. On the other hand, we do not yet know whether $(A \cap R^*) = R^*$ or whether $(A \cap R^*) \subset R^*$, i.e. whether real algebraic numbers comprise all the real numbers there are, or only a (proper) part of them.

The answer is easy: since A is countably infinite, so is $A \cap R^*$ (real algebraic numbers), whereas R^* (all real numbers) is non-countably infinite. Hence, $(A \cap R^*) \subset R^*$, i.e. real algebraic numbers are no more than a proper subset of all real numbers. The situation is indicated in the Venn Diagram of Fig. 6.6.

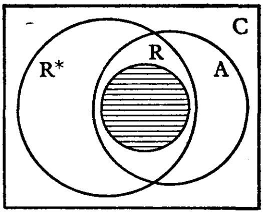
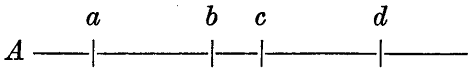


FIG. 6.6

An immediate consequence is that there *are* real numbers which are not roots of polynomial equations, just as there *are* roots of polynomial equations which are not real. Such real numbers are called *transcendentals*. We have shown that they exist. Examples are real numbers like π and e of great importance in mathematics. Moreover, they are not the exception, but very much the rule, among real numbers. Of the non-countably infinite set R^* of all real numbers, the real algebraic numbers are no more than a countably infinite subset.

What is left is the subset of transcendentals, and these are a non-countably infinite set in themselves. Transcendentals are very thick on the ground.

6.7. Ordered fields. The development of real numbers (2.4) depends on the fact that the rationals form, not only a field, but also an ordered field. It is time to put the concept of order on a firmer basis. In a set $\{a, b, c, \ldots\}$ the idea of 'succeeds' or of 'successor' may be taken as undefined. Nothing is added by using the alternative term 'precedes' or 'predecessor': if b succeeds a, then a precedes b. On the other hand, the concept of 'immediate successor', if it is relevant, needs definition. Suppose that (1) b succeeds a, and (2) any other successor c of a is such that c succeeds b. Then b is defined as the immediate successor of a. To visualise, imagine the elements of the set as strung out from left to right, according to their succession, along a line A as in the diagram.



A set $\{a, b, c, \ldots\}$ is **ordered** in this primitive sense if each element (with the possible exception of one, called the last) has an immediate successor. Equally, each element (with the possible exception of one, called the first) has an immediate predecessor. The symbol $<$ is used to indicate order: $a<b$ denotes a is succeeded by b, or a precedes b. The corresponding symbol $>$ is: if $a<b$, then $b>a$. Equality, denoted by $=$, can then be introduced to complete the symbols.

Two illustrations follow:

(i) The set of integers, or any subset, is the case on which this primitive ordering depends. $J=\{\ldots\ -2,\ -1,\ 0,\ 1,\ 2,\ \ldots\}$ is ordered as shown according to $<$, with no first or last term. $J^{+}=\{1, 2, 3, \ldots\}$ is similarly ordered but with a first element. In the present primitive sense, the subset $\{0, 1, 2, \ldots\ n-1\}$ can be taken as ordered, with a first and a last element.

(ii) As something different, consider the order of six houses, all on one side of a road. They can be ordered and numbered from 1 to 6; but there is more than one way of achieving this. The order can be in the spatial sense, houses taken in sequence along the road from one end. But it could be (as in Japan) in a historical or chronological sense, numbering from 1 to 6 being according to the date of completion of building. (The complication of two houses completed at the same date involves the use of $<$ and $=$ in the order.) The second method is not so practical since we move in space and not in time. There are further orderings possible if the six houses are on both sides of the road, e.g. numbering from one end with odd numbers to the left and even to the right, or down one side and back the other.

The definition of an *ordered field** is based on this primitive idea of order, but considerably refined and without reference to immediate successors. In any field, there are sums, negatives and hence differences. The suggestion is that the differences $(b-a)$ of two distinct elements can be positive or negative, which suffices to indicate $a<b$ and $a>b$ respectively. The concept of *order* in a field depends on the (undefined) concept of *positiveness*. This can be formalised:

DEFINITION: *A field $F=\{a, b, c, \dots\}$ is* **ordered** *if it contains positive elements such that:*

(1) *one and only one of the following holds for every element a:*

$$a \text{ positive}, \quad a \text{ zero}, \quad -a \text{ positive}$$

(2) *if a and b are both positive, then $a+b$ and ab are both positive.*

This covers the case of a *positive* element a. By convention, however, where $-a$ is positive, a is said to be a *negative* element of F. The condition (1) is then: a is positive, *or* zero, *or* negative.

Let a and b be any two elements of F. Since F is a field, the difference $b-a$ is defined as an element of F and condition (1) shows that it is positive, *or* zero, *or* negative. Suppose $b-a$ is positive, then we can write $a<b$ and $b>a$. These are alternative and equivalent notations for the same thing: $b-a$ positive. They specify what the symbols '$<$' and '$>$' mean in an ordered field:

NOTATION: *If $b-a$ is positive in an ordered field, write $a<b$ (read: a less than b) and write $b>a$ (read: b greater than a). The notations $a<b$ and $b>a$ are equivalent.*

Suppose $b-a$ is negative. Then $-(b-a)=a-b$ is positive and the new notation gives alternatively: $b<a$ and $a>b$. Finally, suppose $b-a$ is zero. Then $b+(-a)=0$ or b is the negative of $(-a)$. Since a is known to be the negative of $(-a)$, $b-a$ zero implies $a=b$. Condition (1) gives: $b-a$ is positive, *or* zero, *or* negative. This can now

* It can be applied to an integral domain as well as to a field, e.g. J is an ordered domain, but $F[x]$ is not.

be re-written: $a<b$, *or* $a=b$, *or* $a>b$. The condition (2) can be re-expressed in terms of $<$ and $>$ in much the same way, as in 6.9 Ex. 26. So:

THEOREM: *If* $F=\{a, b, c, .\ \}$ *is an ordered field, then*

(i) **Trichotomy:** $a<b$, *or* $a=b$, *or* $a>b$

(ii) **Consistency:** *if* $a<b$, *then* $a+c<b+c$ (*any* c) *and* $ac<bc$ (*any positive* c).

Hence, in an ordered field, we can write $<$, $=$ or $>$ between any two elements. There is no difficulty in using $\leqslant$ for '$<$ or $=$', and $\geqslant$ for '$>$ or $=$'. In particular, since 0 is an element of the field, $a>0$ can be written for $a-0$ positive, i.e. a positive; and $a<0$ can be written for $0-a$ positive, i.e. $-a$ positive, a negative.

The field R of rationals is ordered since positiveness is ensured by the convention that the rational $p/q>0$ if the product of integers $pq>0$. The trichotomy and consistency properties of the theorem above are among the order properties of rationals as set out in 2.2.

On the other hand, examples of non-ordered fields are easily got. Neither the field C of complex numbers nor the field $F(x)$ of rational fractions is ordered, since the elements do not have the trichotomy of positive/zero/negative. It is not so obvious that the field of integers (mod prime n) is not ordered in the present precise sense. It is condition (2) which fails: if a and b are positive, then $a+b$ is positive. Here, 1 and 2 are positive, but $1+2=0$ is not, in the field of integers (mod 3).

The concept of an ordered field can be developed as in 2.4:

DEFINITION: *A field* F *is* **completely ordered** *if it is ordered and such that: any subset of* F *which has a lower bound has a GLB; any subset of* F *which has an upper bound has a LUB.*

In the number system only the field R^* of real numbers is completely ordered.

6.8. Inequalities. The properties of an ordered field provide all the material needed for handling inequalities. There is a relation of *equality* in *any* field. For example, in the non-ordered field of complex numbers, one number can be equated to another, involving the process of 'equating real and imaginary parts'. But the relation of *inequality* arises when we have an *ordered* field. Inequalities are subject to operational rules, derived from the properties of ordered fields; since they are somewhat tricky, we must lay them out carefully.

In the following properties, $a<b$ (or $b>a$) means $b-a$ positive in an ordered field $F=\{a, b, c, \dots\}$:

(i) If $a<b$, then $a+c<b+c$ (any c) and $ac<bc$ $(c>0)$ (Consistency).
(ii) If $a<b$ and $b<c$, then $a<c$ (Transitivity).
(iii) If $a<b$ and $c<d$, then $a+c<b+d$.
(iv) If $a<b$, then $-a>-b$.
(v) If $a<b$ and $ab>0$, then $1/a>1/b$.

The further properties below relate to the particular case of positive and negative elements, i.e. $a>0$ means a positive, $a<0$ means a negative:

(vi) If $ab>0$, then *either* $a>0$, $b>0$ *or* $a<0$, $b<0$.
If $ab<0$, then *either* $a>0$, $b<0$ *or* $a<0$, $b>0$.
(vii) If $a>0$, then $1/a>0$. If $a<0$, then $1/a<0$.
(viii) If $a>b>0$, then $1/b>1/a>0$. If $a<b<0$, then $1/b<1/a<0$.

Many of the proofs are straightforward applications of the definition of ordered fields in 6.7. One or two are, however, a little awkward. Notice, for example, property (v). Here, $a<b$ does not necessarily imply $1/a>1/b$. This implication is valid if and only if a and b have the same 'sign', meaning $ab>0$ as indicated in property (vi). For example, $2<3$ does imply $1/2>1/3$ and $-3<-2$ does imply $-1/3>-1/2$; but, though $-2<3$, it is not true that $-1/2>1/3$.

A characteristic feature of an ordered field F is that the square of any non-zero element is necessarily positive:

$$a^2=0 \text{ if } a=0;\ a^2>0 \text{ if } a\neq 0.$$

This is not necessarily so in a non-ordered field. For example, in the field of complex numbers, the element i has a real square i^2 but it is negative: $i^2=-1$. The result extends:

If $a_1, a_2, \dots a_n$ are elements of an ordered field, then

$$\sum_{r=1}^{n} a_r{}^2>0 \quad \text{unless } a_1=a_2=\dots=a_n=0.$$

To illustrate the handling of equalities and inequalities, consider cases of polynomials or rational fractions where x is taken as a real number. Pursuing the distinctions of 1.4 and 1.5 above, we say that an equation holds for certain x which compose its *solution set*, as in the simple examples:

	Equation	*Solution set*
(i)	$5-2x=0$	$x=\frac{5}{2}$
(ii)	$x^2-3x+2=0$	$x=1, 2$
(iii)	$x^2-3x+2=\frac{6}{x}$	$x=3$

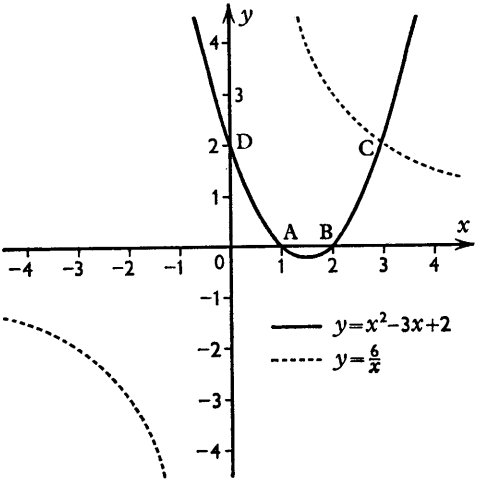


Fig. 6.8

Graphically, the solution set corresponds to the intersection of two sets, each consisting of pairs of real numbers (x, y) shown as points in the plane Oxy. For example, in (ii), $y=x^3-3x+2$ gives the set of points (x, y) of the curve shown in Fig. 6.8. If the second set is $y=0$, i.e. the line Ox, then the solution set for (ii) is the intersection of the two sets. This is the pair of points $A(1, 0)$ and $B(2, 0)$. So: $x=1, 2$. In (iii), with the same first set, then $y=6/x$ gives another set of points (x, y) on the second curve of Fig. 6.8. The intersection set is now the single point $C(3, 2)$. So: $x=3$.

Now consider inequalities in the same way:

Inequality	*Solution set*
(i) $5-2x>0$	$x<\frac{5}{2}$
(ii) $x^2-3x+2<0$	$1<x<2$
(iii) $x^2-3x+2<\frac{6}{x}$	$0<x<3$

To spell these out carefully, we proceed as follows. For the first two inequalities:

(i) $5-2x>0$ gives $-2x>-5$ by property (i) with $c=-5$
gives $2x<5$ by property (iv)
gives $x<5/2$ by property (i) with $c=1/2>0$

(ii) $x^2-3x+2=(x-1)(x-2)<0$ gives by property (vi):
either $x-1>0$ and $x-2<0$
i.e. $x>1$ and $x<2$ by property (i)
or $x-1<0$ and $x-2>0$
i.e. $x<1$ and $x>2$ by property (i).

The first is consistent, the second not. Hence $1<x<2$.

Graphical methods again give the solution set as the intersection of two sets. In (ii), the set of points on the curve $y=x^2-3x+2$ and the set of points ($y<0$) below the axis Ox intersect in the set of points on the curve from A to B. For these points: $1<x<2$. In (iii), the same curve is related to the set of points ($y<6/x$) below the second curve $y=6/x$. The intersection set consists of points on the first curve between D and C. For these points: $0<x<3$.

6.9. Exercises

1. A group depends on a binary operation $*$, usually $+$ or $\times$; explain why neither subtraction ($-$) nor division ($\div$) is a suitable operation $*$. See 2.9 Ex. 3.

2. Show that $S=\{\dots -4, -2, 0, 2, 4, \dots\}$ is a group under $+$ and hence a subgroup of the group J of all integers. Check by the conditions (6.2) for a subgroup. Extend to the set of multiples of any integer.

3. Show that the set $\{\dots \frac{1}{4}, \frac{1}{2}, 1, 2, 4, \dots\}$ of integral powers of 2 is a group under $\times$, a subgroup of the group R of all rationals. Extend as in Ex. 2.

4. Show that the set R^* of real numbers is a group under $+$ and also (when 0 is excluded) under $\times$; establish the same result for the set C of complex numbers.

5. *Solution of linear equations in a group.* $G = \{a, b, c, \ldots\}$ is a group under $*$, so that $x = b^{-1} * a$ is an element of G. Show that x satisfies $b * x = a$. Further, if x_1 and x_2 both satisfy $b * x = a$, show that $x_1 = x_2$. Deduce that the linear equation $b * x = a$ has a unique solution $x = b^{-1} * a$ in a group.

6. As a case of Ex. 5, show that an additive group (commutative) is characterised by the fact that $b + x = a$ and $x + b = a$ alike have the unique solution $x = a - b$.

7. Show that, in a group under ×, the linear equation $bx = a$ has the unique solution $x = b^{-1}a$, and $xb = a$ the unique solution $x = ab^{-1}$. If the group is commutative, $bx = a$ and $xb = a$ equally have the unique solution $x = \frac{a}{b}$.

8. *Cyclic groups under addition.* If $na = 0$, show that $\{a, 2a, 3a, \ldots na\}$ is a finite cyclic group under $+$ with group identity 0. Illustrate by writing the integers (mod n) as $\{1, 2, 3, \ldots n\}$ where $n = 0$.

9. Show that the × table for integers (mod n) has 0 in the pth row and qth column if and only if $n = pq/r$ for some positive integer r. Indicate how this supports the proposition that the integers (mod n), excluding 0, form a group under × if and only if n is prime.

10. In a group under $*$, show that the inverse of $a * b$ is $b^{-1} * a^{-1}$ (and not $a^{-1} * b^{-1}$ unless the group is commutative). Start from $b^{-1} * a^{-1} * a * b$ and reduce by using $a^{-1} * a = b^{-1} * b = e$.

11. In a group with an identity e such that $a * e = e * a = a$ (and with cancellation valid), deduce that e is unique. (Show that, if there are two e_1 and e_2, so that $a * e_1 = a * e_2 = a$, then $e_1 = e_2$.) Similarly, show that a^{-1} such that $a * a^{-1} = a^{-1} * a = e$ is unique.

*12. *Cosets.* Partition the set J of all integers into the set J_0 of even integers and the set J_1 of odd integers. Under $+$, the group J has subgroup J_0 (see Ex. 2). Check that J_1, which is not a group, is obtained from J_0 by adding 1 to each element. Sets J_0 and J_1 are called *cosets*. Extend the result, e.g. to the case of five cosets of J: all multiples of 5 making up the subgroup J_0 and J_r ($r = 1, 2, 3, 4$) being obtained by adding r to each element of J_0. Compare 3.9 Ex. 18 on residue classes. (The sets J_r are both residue classes — when derived from remainders on division by a given integer — and also cosets — when obtained by adding the same integer to the elements of some subgroup J_0.)

*13. *Factor group.* Consider the set $\{J_0, J_1\}$ of two elements, the sets defined in Ex. 12. Define $J_r + J_s$ = set of elements obtained by adding an element of J_r to an element of J_s ($r = 0, 1$; $s = 0, 1$). Show that $J_0 + J_0 = J_1 + J_1 = J_0$ and that $J_0 + J_1 = J_1 + J_0 = J_1$. Hence show that $\{J_0, J_1\}$ is a group under $+$. Such a group is called a *factor group*. Extend as in Ex. 12.

*14. Partition $S = \{i, -1, -i, 1\}$, a cyclic group under ×, into $S_0 = \{1, -1\}$ and $S_1 = \{i, -i\}$. Show that S_0 is a group but not S_1. Each element of S_1 is a multiple i of an element of S_0, another example of cosets. Show that the set $\{S_0, S_1\}$ of two elements is a group under ×, with the definition that $S_r \times S_s$ = set of elements obtained by forming products of an element of S_r and an element of S_s ($r = 0, 1$; $s = 0, 1$). This is another factor group.

15. Rotate axes Oxy through $\theta°$ (clockwise) to give new axes $Ox'y'$ as in Fig. 6.9. The fixed point P has co-ordinates (x, y) changed to (x', y'). By showing that $OM' = OM \cos\theta - MP \sin\theta$ in the figure, establish that

$$x' = x \cos\theta - y \sin\theta$$

and $$y' = x \sin\theta + y \cos\theta$$

similarly. Show that this is also the transformation for the rotation of a figure through $\theta°$ (anti-clockwise), axes being fixed.

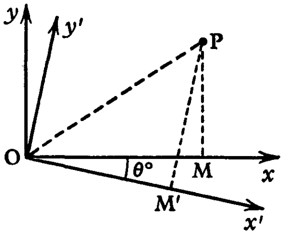


FIG. 6.9

16. *Movements of a square.* First rotate a square through 0°, 90°, 180° and 270° and show that the transformations can be denoted by $1, i, -1, -i$. Then turn the square over (i) horizontally, (ii) vertically, (iii) about the diagonal sloping upwards from left to right, (iv) about the diagonal sloping downwards from left to right. Write these transformations p, q, r and s. Completing the table of pairs of transformations as shown, deduce that the transformations form a non-commutative group.

×		*First transformation*							
		1	i	-1	$-i$	p	q	r	s
Second:	1	1	i	-1	$-i$	p	q	r	s
	i	i	-1	$-i$	1	r	s	q	p
	-1	-1	$-i$	1	i	q	p	s	r
	$-i$	$-i$	1	i	-1	s	r	p	q
	p	p	s	q	r	1	-1	$-i$	i
	q	q	r	p	s	-1	1	i	$-i$
	r	r	p	s	q	i	$-i$	1	-1
	s	s	q	r	p	$-i$	i	-1	1

17. In the group of Ex. 16, show that the subset of transformations $\{1, i, -1, -i\}$ form a cyclic subgroup which is commutative.

18. From the + and × tables for $\{0, 1, 2, 3, 4\}$ (mod 5), check that the distributive rule $a(b+c) = ab + ac$ holds, by trying out various a, b and c.

19. Show that $\{-1, 0, 1\}$ is a field, i.e. a commutative group under + and (excluding 0) a commutative group under ×, if $1 + 1 = 0$ is imposed. Check that the distributive rule holds.

20. The additive identity (zero, 0) of any field F is excluded from the multiplicative group. But 0 can be used in multiplication: $a \times 0 = 0$ for any a. Consider the question whether $a \times 0 = 0$ should be specified as a requirement for a field, in linking + and ×. Show that the answer is no, since $a \times 0 = 0$ follows from the distributive rule. (Start with $a(1+0) = a \times 1 + a \times 0$, write $1 + 0 = 1$ and use the additive cancellation rule.)

21. *Subfields.* Show that necessary and sufficient conditions for a subset K to be a subfield of a field F are that K contains at least one non-zero element and that, if $a \in K$, $b \in K$, then $a - b \in K$ and $ab^{-1} \in K$ where $b \neq 0$.

*22. *Quaternions.* Extend the concept of a complex number to the quaternion, defined $z=a+ib+jc+kd$ where a, b, c and d are real and where products are subject to the conventions shown on products of i, j and k. Show that the set of z's (all real a, b, c and d) satisfies the rules for a field, except that products are not commutative. Such a set is called a *skew field.*

×	*First element:*		
	i	j	k
Second: i	-1	$-k$	j
j	k	-1	$-i$
k	$-j$	i	-1

*23. *Field extensions.* Pull together various ways of extending a given field F into a larger field. First, from F is obtained the integral domain $F[x]$ of polynomials over F and the quotient field formed to turn $F[x]$ into the field $F(x)$. This is the field of rational fractions, ratios of polynomials over F (3.4). Second, the adjunction of an outside element x to F gives a field $F(x)$, which is also to be identified with the set of ratios of polynomials over F (3.4). The field of complex numbers is obtained from the field of real numbers by adjunction of i in this way. Third, select from the integral domain $F[x]$ a polynomial $g(x)$ irreducible in F and write the field of polynomials mod $g(x)$ (see 3.9 Ex. 20). The field of complex numbers arises from the field of real numbers by taking $g(x)=x^2+1$ in this way (and see 7.9 Ex. 20).

*24. The field $F=\{0, 1\}$ (mod 2) contains two elements only. The set of polynomials over F can be enumerated (see 3.9 Ex. 11):

$$0,\ 1,\ x,\ x+1,\ x^2,\ x^2+1,\ x^2+x,\ x^2+x+1,\ \ldots$$

The polynomial x^2+x+1 is irreducible and the set of these polynomials, mod x^2+x+1, is a field. Show that it consists of four elements: $\{0, 1, x, x+1\}$. Check that, in this field, x and $x+1$ add and multiply alike to 1. [Note: $x+(x+1)=2x+1=1$ (mod 2); $x(x+1)=x^2+x=-1$ (mod x^2+x+1) $=1$ (mod 2).]

25. Illustrate that a set which is not an ordered field can have a subset which is an ordered field by reference to the set C of all complex numbers.

26. If $a<b$, show that $a+c<b+c$ (any c) and $ac<bc$ (any $c>0$), by use only of the criterion that $a<b$ if $b-a$ is positive.

27. Define an ordered integral domain (lacking only reciprocals) as for an ordered field and illustrate with the set J of integers.

28. Show that, if $a<b$ and $c<d$, then $a+c<b+d$, but it is not necessarily the case that $ac<bd$. Illustrate by examples in which a and c are both negative while b and d are of opposite sign.

29. Prove and illustrate that $\dfrac{1}{b}>\dfrac{1}{a}>0$ if $a>b>0$ and $\dfrac{1}{b}<\dfrac{1}{a}<0$ if $a<b<0$.

30. Show that the inequalities $x^2+x+1\leqslant 1/(1+x^2)$ and $x>0$ have a solution set which is empty. Deduce that $x^2+x+1>1/(1+x^2)$ for all $x>0$.

CHAPTER 7

RELATIONS AND FUNCTIONS

7.1. Relations. Consider sets of elements in the general sense of Chapter 4; they may be groups or fields, but they may equally have some quite different structure. Let X and Y be two sets and write $x \in X$ and $y \in Y$ as representative elements. Subscripts can then be used to indicate various elements: x_1, x_2, x_3, ... as elements of X; y_1, y_2, y_3, ... as elements of Y.

Form the set of *ordered pairs* (x, y), ordered in the sense of x (from X) first and y (from Y) second and not the other way round.* To illustrate in graphical terms, suppose that X and Y are sets of integers, rationals or real numbers. Then the elements of X can be marked off on one axis Ox and the elements of Y on another axis Oy, the ordered pair (x, y) being shown by a point in the plane Oxy. If it happens that Y contains zero (0), then particular points $(x, 0)$ on Ox are in the set of (x, y) and correspond to the set X as well. Similarly for the set Y on Oy. As a notation, the set of (x, y) is called the Cartesian product and written $X\,.\,Y$:

NOTATION: *The* **Cartesian product** *of two sets X and Y is the set of ordered pairs*

$$X\,.\,Y = \{(x, y) \mid x \in X, y \in Y\}.$$

Since the order of writing x and y is essential, the Cartesian products $X\,.\,Y$ and $Y\,.\,X$ are different. As a particular case, for the single set $S = \{x, y, z, \ldots\}$, there is a Cartesian product: $S\,.\,S = \{(x, y) \mid x \in S, y \in S\}$. This is *not* the same as the smaller set $\{(x, x) \mid x \in S\}$. As simple examples, take $X = \{1, 2\}$ and $Y = \{1, 2, 3, 4\}$, so that $X\,.\,Y$ is the set of 8 elements illustrated by 8 points in Fig. 7.1*a*. Similarly $Y\,.\,X$ is a different set of 8 points; $X\,.\,X$ is a set of 4 points and $Y\,.\,Y$ one of 16 points.

* Strictly, an ordered pair is a primitive (undefined) concept subject to an axiom: the property that $(x, y) = (u, v)$ implies $x = u$ and $y = v$.

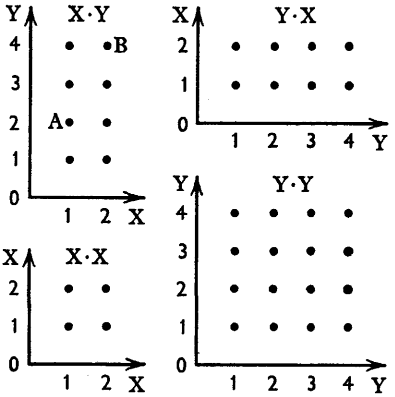


FIG. 7.1*a*

The concept of a *relation* is a broad and sweeping one. Any proper subset of $X \, . \, Y$ is a relation R from the set X to the set Y. Hence, R is a set: the set consisting of some but not all of the ordered pairs (x, y) of $X \, . \, Y$. R is to be distinguished from its specification, which may be by listing or by a general description of the pairs (x, y) in R. The specification is written yRx, to be read 'y is related by R to x', giving the rule for connecting x in X and y in Y. R is a *set* and the specification yRx is a *statement* or a *rule*. The concepts of Chapter 5 are relevant since we have both a statement yRx and a corresponding set R. In the following, implication (p implies q, or if p then q) and equivalence (p and q equivalent, or if p then q *and* if q then p) are freely used as statements true in all logical possibilities.

DEFINITION: *A* **relation** *R from the set X to the set Y is any proper subset of X . Y. The specification yRx implies* $(x, y) \in R$ *so that:*

$$R = \{(x, y) \mid x \in X, y \in Y, yRx\}.$$

The statement yRx is to be read 'y is related by R to x'.

Each $x \in X$ may or may not have a $y \in Y$ to correspond in yRx. Those x for which there *are* y's make up a subset of X called the **domain** of R. Similarly, each $y \in Y$ may or may not correspond to an $x \in X$ in yRx. Those y for which there *are* x's make up a subset of Y called the **range** of R.

In speaking of a relation, we can use the set R and the statement yRx interchangeably. Some examples illustrate:

(i) $X = \{1, 2\}$, $Y = \{1, 2, 3, 4\}$. Write:

$$R_1 = \{(x, y) \mid x \in X, y \in Y, y = 2x\} \text{ and } R_2 = \{(x, y) \mid x \in X, y \in Y, y < 2x\}.$$

The set of 8 elements $X \, . \, Y$ has a subset of 2 elements, (1, 2) and (2, 4), for R_1, i.e. the points A and B in Fig. 7.1*a*. It has a subset of 4 elements, (1, 1), (2, 1), (2, 2) and (2, 3), for R_2, i.e. the points below

A and B in Fig. 7.1a. The statements are: yR_1x for '$y=2x$' and yR_2x for '$y<2x$'.

(ii) X and Y both comprise all real numbers. Specify R_1 and R_2 as in (i). Then R_1 given by the statement '$y=2x$' is the set of points on the line $y=2x$ in the plane Oxy, and R_2 for '$y<2x$' is the set of points under the line $y=2x$. They are both subsets of $X \,.\, Y$, the set of all points in the plane.

(iii) X and Y both comprise all positive real numbers ($x>0$, $y>0$); $X \,.\, Y$ is the set of points in the positive quadrant of the plane Oxy. Consider the relations:

$$R_1=\{(x, y) \mid x\epsilon X, y\epsilon Y, x+y=1\};$$
$$R_2=\{(x, y) \mid x\epsilon X, y\epsilon Y, x+y<1\};$$
$$R_3=\{(x, y) \mid x\epsilon X, y\epsilon Y, x^2+y^2=1\};$$
$$R_4=\{(x, y) \mid x\epsilon X, y\epsilon Y, x^2+y^2<1\}.$$

The sets R_1 and R_3 are shown by the points on the segment AB of the line $x+y=1$, and on the arc AB of the circle $x^2+y^2=1$, respectively in Fig. 7.1b. The set R_2 is shown by the points within the triangle OAB, as shaded; the set R_4 corresponds to points within the quarter-circle OAB.

(iv) Consider the 16 members of the tribe of 4.1, example (iii), and write X for the set of 7 males and Y for the set of 9 females. Then $X \,.\, Y$ is the set of 63 different pairings of male/female. Define the relation:

$R=\{(x, y) \mid x \in X, y \in Y, y$ is the wife of $x\}$.

Here yRx is the statement 'y is the wife of x'. The set R comprises 4 out of 63 pairs, two married couples in the first and two in the second generation.

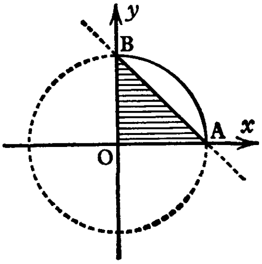


FIG. 7.1b

The variety of relations is very great: *any limitation however slight on x and y gives a relation.* The x's and y's may be numbers, discrete as in (i) or real variables as in (ii) and (iii); the relations may involve an equation or equally well an inequality. But relations can be perfectly well specified between non-numerical x's and y's, as in (iv) where the elements are persons and the relation is the ordinary one of wife to husband.

7.2. Equivalence. Consider first an important, if very special, kind of relation, that of equivalence between elements of a given set

$S=\{x, y, z, \dots\}$. In a set of numbers it means equality ($=$); in other sets, it has a wider connotation. For example, let S consist of all finite sets $A, B, C, \dots$ as in 4.5. Then equivalent sets are those which have the same count of elements. Equivalence, in fact, is the concept of things being the 'same' or 'alike' in some way.

DEFINITION: *yRx is an* **equivalence relation** *in a set S if the properties:*

(1) **Reflexive:** *xRx logically true;*
(2) **Symmetric:** *if yRx, then xRy;*
(3) **Transitive:** *if zRy and yRx, then zRx;*
hold for any x, y and z in S.

The emphasis here is on the statement yRx of equivalence, rather than on the set R of pairs (x, y) of equivalent elements of S within the set $S \,.\, S$ of ordered pairs. However, the set R is always to be remembered. A feature of equivalence is that R is symmetrical, i.e. the ordering in this special case does not matter.

For example, let S be the set of positive rationals $\frac{p}{q}$ (p and q positive integers). Then $\frac{r}{s}R\frac{p}{q}$ is an equivalent relation given by 'the integers ps and qr are equal'. It can be shown to satisfy all three properties. Having established the equivalence, we usually write R as $=$, i.e. $\frac{r}{s}=\frac{p}{q}$ if $ps=qr$. Further, fixing one rational $\frac{p}{q}$, we can find all the rationals $\frac{r}{s}$ equivalent (equal) to it. If $\frac{1}{2}$ is fixed, the equivalent rationals in S are $\frac{1}{2}, \frac{2}{4}, \frac{3}{6}, \dots$. As a set, R can be written

$$R = \left\{\left(\frac{p}{q}, \frac{r}{s}\right) \middle| \frac{p}{q} \in S, \frac{r}{s} \in S, ps = qr\right\}.$$

The concept of equivalence is, however, far wider. It is one of the basic ideas in logic and mathematics. Consider the set S of all finite sets $A, B, C, \dots$. Take the relation R of 'equi-numerous' as defined in 4.5. Then the following all hold: ARA; if ARB, then BRA; if ARB and BRC, then ARC. Write them with R replaced by $\sim$ for 'equi-numerous': $A \sim A$; if $A \sim B$, then $B \sim A$; if $A \sim B$ and $B \sim C$, then $A \sim C$. From the counting aspect, equi-numerous sets are alike; they are equivalent.

One thought may now occur to us. We have used the term 'equivalence' for statements p and q in the sense: if p then q *and* if q then p. This is a relation R between p and q. Is it an equivalent relation, so that our terminology is consistent? It is easily checked to be so, by taking the three properties in turn:

Reflexive: pRp i.e. 'if p then p and if p then p', which is true.

Symmetric: pRq and qRp are the same i.e. 'if p then q and if q then p' is the same as 'if q then p and if p then q', which is so.

Transitive: from pRq and qRr we get pRr, i.e. 'if p then q and if q then p' taken with 'if q then r and if r then q' gives 'if p then r and if r then p', which is the case.

Carrying out checking of this nature is not often necessary, since it is reasonably obvious in most cases of equivalence that the properties are valid.

Equivalence has to do with the *partitioning* of a given set S. Of the various subsets of S which can be written, many are overlapping. Even if they are disjoint, they may or may not exhaust S between them. A partition of S is a set of subsets which are both disjoint and exhaustive of S. As an example, consider the set of 100 people classified according to their smoking habits, as given in 4.6. The seven subsets given in the original data overlap and do not exhaust the whole set of 100. The job undertaken in 4.6 is to get the numbers of elements common to the subsets and to derive the 'remainder' (the non-smokers). Partitions of the set S of 100 people can then be written in various ways. With the notation: A for cigarette smokers, B for cigar smokers, C for pipe smokers, one partition of S is:

$$S = A + A'B + A'B'C + A'B'C' \quad (100 = 42 + 10 + 10 + 38)$$

i.e. S is all cigarette smokers + cigar smokers not smoking cigarettes + pipe smokers not smoking cigarettes or cigars + non-smokers. A similar partition is:

$$S = B + B'C + AB'C' + A'B'C' \quad (100 = 17 + 19 + 26 + 38).$$

Another partition is of a particular kind, a finer grouping of the first one:

$$S = \underbrace{ABC + ABC' + AB'C + AB'C'}_{A} + \underbrace{A'BC + A'BC'}_{A'B} + A'B'C + A'B'C',$$

This finer grouping, a partitioning of S into 8 subsets, is the one displayed in the Venn Diagram of 4.6.

The connection between equivalence and partitioning follows from some simple propositions on an equivalence relation yRx in the (non-empty) set $S=\{x, y, z, \ldots\}$:

(*a*) If S_x is the set of y equivalent to a given $x\,(yRx)$, then at least one (non-empty) S_x exists as a subset of S. Since S is not empty, it contains an element x and $x \in S_x\,(xRx)$.

(*b*) If yRx, then $S_x=S_y$ and conversely. The direct part is established: if yRx is given, then xRy also. Let $z \in S_x$, so zRx which with xRy gives zRy, i.e. $z \in S_y$. Hence $S_x \subseteq S_y$. This same kind of argument gives $S_y \subseteq S_x$. So: $S_x=S_y$. The converse: given $S_x=S_y$, it follows that $y \in S_y$ (since yRy) and hence that $y \in S_x$. Hence yRx by the definition of S_x.

(*c*) If $S_x \neq S_y$ exist, then they are disjoint. Suppose S_x and S_y have a common element z, so that zRx and zRy. But zRy gives yRz, which (with zRx) gives yRx. By (*b*), $S_x=S_y$, a contradiction. Hence S_x and S_y are disjoint. Q.E.D.

Now consider all the sets like S_x which can be written and which are distinct. There is at least one by (*a*). If more than one, they are disjoint by (*c*). By (*b*), if x and y are equivalent (yRx) they go into the same S_x; otherwise, into distinct S_x and S_y. Finally, all the sets like S_x exhaust S since, if there were an element z left over, then S_z could be formed with z in it (zRz) and added to the sets already written. Hence:

THEOREM: *An equivalence relation yRx in a set S determines a partition of S into disjoint and exhaustive subsets S_x. Two elements x and y of S are in the same S_x if and only if yRx.*

The subsets S_x are called *equivalence classes* of S. All elements of S which are equivalent to a given element x (and equivalent to each other) go into one and the same S_x. An element x of S_x can be taken as representative of all, and called the *canonical form* of S_x. All other elements of S_x are equivalent to the canonical form x. Note that the partition may consist of only one equivalence class (if all elements of S are equivalent), and that it may include equivalence classes which have only one member. Some examples follow:

(i) $S=\{A, B, C, \ldots\}$ comprising finite sets. The equivalence relation is $A \sim B$, or 'A has the same number of elements as B'. The partition of S is into equivalence classes S_1, S_2, S_3, ... where S_r comprises all sets with r elements.

(ii) S is the set of 100 people classified by smoking habits (4.6). The equivalence relation is 'x and y have the same smoking habits', meaning that x and y are alike in smoking cigarettes (or not), cigars (or not), pipe (or not). The partition of S is into 8 subsets as shown above. The first subset ABC consists of all those equivalent people who smoke all three; and similarly for the others.

(iii) S is the set of 16 tribal members of 4.1, example (iii). The equivalence relation is yRx for 'x and y are of the same generation'. S is partitioned by generations into an equivalence class of 4 people (first generation), another of 8 people (second generation) and a third of 4 people (third generation).

(iv) The group J of integers under addition. The equivalence relation is 'x and y have the same remainder on division by 3'. The partition of J is:

Subset	*Canonical form*
$J_0 = \{\ldots \quad -3 \quad 0 \quad 3 \quad 6 \quad \ldots\}$	0
$J_1 = \{\ldots \quad -2 \quad 1 \quad 4 \quad 7 \quad \ldots\}$	1
$J_2 = \{\ldots \quad -1 \quad 2 \quad 5 \quad 8 \quad \ldots\}$	2

This is linked with the group of integers (mod 3) under addition. The set of equivalence classes is $\{J_0, J_1, J_2\}$; the set of canonical forms is $\{0, 1, 2\}$ which is the group of integers (mod 3).

(v) The cyclic group $S = \{i, i^2, i^3, i^4\} = \{i, -1, -i, 1\}$ under multiplication. The absolute value of a complex number

$$|a + ib| = \surd(a^2 + b^2),$$

giving the equivalence relation: 'x and y have the same absolute value'. Then S is itself one equivalence class.

7.3. Functions and mappings. A relation yRx from the set X to the set Y is such a general concept that it does not imply that, for each x in the domain of X, there is just one y in the range of Y; there may well be several such y's. For example, if x and y are real numbers, the relation $x + y = 1$ does give one y for each x; but $y^2 = x$ gives no y when x is negative, one y (i.e. $y = 0$) when $x = 0$ and two y's ($y = \pm \surd x$)

when x is positive; and $x+y<1$ is a relation in which (infinitely) many y's correspond to each x.

A relation f such that just one y corresponds to each x in the domain is of the greatest importance; it is called a *functional relation* or more simply a *function*. The term 'function' can be used, when there is no ambiguity, to indicate both the set of the functional relation f and the statement yfx (see 9.1 below). So, yfx can be read 'y is a function of x' and replaced by the more usual $y=f(x)$.

DEFINITION: *A* **function** *f from the set X to the set Y is a relation specified by the statement yfx such that, if $y \in Y$ exists for a given $x \in X$, then y is unique. The statement yfx can be written $y=f(x)$, read 'y is a function of x'.*

In full, in terms of sets, f is the set $\{(x, y) \mid x \in X, y \in Y, y=f(x)\}$. As a further essential notation, the function f is defined on the **domain** $\{x \mid \text{there exists } y \text{ such that } y=f(x)\}$, a subset of X; and the **range** of f is $\{y \mid \text{there exists } x \text{ such that } y=f(x)\}$, a subset of Y.

Usually, with little loss of generality, the set X can be limited to the domain of f. We say then: $y=f(x)$ defined on X. The range is still a subset of Y consisting of y's such that $y=f(x)$ for some x in X. There is a unique y in the range for each specified x and the function $y=f(x)$ is often called *single-valued*.* The converse is *not* generally true; for each specified y in the range of $y=f(x)$, there may well be more than one x to correspond. Some examples illustrate:

(i) $y=2x$ defined on the domain X of *all* positive integers. This gives one y for each x (as required) and the range is the set of *even* positive integers. It also gives one x (i.e. $x=\frac{1}{2}y$) for each y in the range.

(ii) $y=x^2$ defined on the domain X of *all* real numbers. This gives one y for each x (as required) and the range is the set Y of *non-negative* real numbers. For each $y \neq 0$ in the range, there are two x's to correspond ($x= \pm \sqrt{y}$). This function also becomes single-valued both ways if it is defined on the domain X of all *non-negative* real numbers, which is also the range ($y=x^2$, $x=+\sqrt{y}$).

(iii) A function need not be algebraic or analytical. It can very well be non-analytical. So: $y=f(x)$ where 'y is the wife of x' is a perfectly respectable function under monogamy. This is so in the tribe of 7.1, example (iv).

* The definition here does *not* cover what is called a multi-valued function (e.g. $x^2+y^2=1$, x a real number). These need to be split into single-valued branches.

A function can be viewed from a rather different angle if X and Y comprise real numbers. A relation, as a subset of the Cartesian product $X \,.\, Y$, is shown as a subset of points in the plane Oxy. A function is represented by a particular kind of subset of points; all lines parallel to Oy (and corresponding to the domain of X) intersect the subset of points in just a single point. The representation of a function is a more or less recognisable 'curve' proceeding from left to right over the domain of X. This is the *graphical* aspect of a function.

An equally important diagrammatic aspect is that of a function as a *mapping*. This merits a separate definition:

DEFINITION: *The function $y=f(x)$, defined on X and with a subset of Y as range, gives a* **mapping** *of the set X into the set Y, denoted $X \underset{f}{\rightarrow} Y$ such that to each $x \in X$ there is a unique* **image** $f(x) \in Y$.

It must be stressed that a mapping is the same thing as a function; nothing new is introduced. At the same time, a mapping is often a good way of looking at a function, a very convenient expression of the relation involved. Functions, though definable for sets of any kind, are so often used for numbers (numerical variables) that it is difficult to avoid thinking that the association is inevitable. There can be no such inhibition for mappings. A set X of any kind can be mapped into another set Y of the same or different kind. For example, the mapping $y=2x$ of the set X of positive integers $\{1, 2, 3, \ldots\}$, into the set Y of positive integers, can be shown as in Fig. 7.3. Each element of X has its unique image in the set Y. Exactly the same diagram might serve (for example) for the mapping $y =$ wife of x, where X is the set of married men in a tribe and Y the set of women. Other kinds of diagrams can be drawn for mappings (e.g.) of a set of points in three dimensions into a set of points in two dimensions. In a mapping, all we require of $f(x)$ is that it lays down a *definite rule* for getting from x in X to its unique image y in Y.

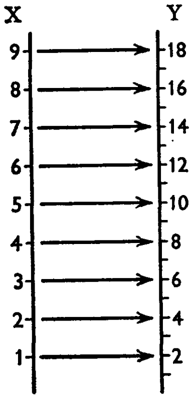


FIG. 7.3

The following distinctions can usefully be made about a mapping f and about the corresponding function $y=f(x)$:

(*a*) The whole of the set X is here taken as the domain so that each

$x \in X$ has its unique image $y \in Y$. The converse is not necessarily true; two or more $x \in X$ can correspond to a single specified image $y \in Y$. The mapping is generally *many-one*. It is a special case when the mapping is *one-one*.

(*b*) The sets X and Y are generally different. It is a special case when X is mapped *into itself*.

(*c*) The range of $y=f(x)$ is a subset of the set Y; generally it will be a *proper* subset and the mapping is of X *into* Y. It is a special case when the range is the whole of Y and the mapping is of X *onto* Y.

These distinctions can be illustrated:

(iv) $y=2x$ is a mapping of $X=\{1, 2, 3, \ldots\}$ into itself, as illustrated. The mapping is both *one-one* and *into*. The range of y is a proper subset of X. Again, $y=x^2$ is a mapping of the set X of all real numbers *into* itself, or *onto* the set Y of non-negative real numbers. In each case, the mapping is *two-one*.

(v) Consider the set X of three objects $\{A, B, C\}$ and the permutation: A to C, B to A, C to B. This is a one-one mapping f of X *onto* itself, the rule for getting images being:

$$f(A)=C, \quad f(B)=A, \quad f(C)=B.$$

Any mapping of a finite set onto itself must be one-one (and so a permutation).

(vi) X is the set of (currently) married men, and Y the set of all women, in a tribe. Consider the relation 'y is the wife of x'. In a monogamous tribe, this is a *one-one* mapping of X *into* Y. If the conjugal convention is polyandry, it is a *many-one* mapping of X *into* Y. Under polygamy, however, the relation is not a function and there is no mapping; a man's image (wife) is then not unique.

Consider the concept of one-one correspondence, defined in 4.5. In a general sense, a correspondence between two sets X and Y is many-many. It is only useful, however, if it is many-one, in which case it is a function or mapping. The particular case of a one-one correspondence is that of a one-one mapping. The notation for mappings can usefully be extended:

(*a*) If $y=f(x)$ is a many-one mapping of X into Y, write $X \xrightarrow[f]{} Y$, and for particular images $y_1=f(x_1)$, $y_2=f(x_2)$, ... write $x_1 \rightarrow y_1$, $x_2 \rightarrow y_2$, ...

(*b*) If $y=f(x)$ is a one-one mapping of X into Y, write $X\underset{f}{\leftrightarrow}Y$, and for particular images $y_1=f(x_1)$, $y_2=f(x_2)$, ... write $x_1\leftrightarrow y_1$, $x_2\leftrightarrow y_2$, ...

As a final example of a many-one correspondence (mapping), consider the link between letters and digits on the dial of a London telephone instrument:

	ABC	DEF	GHI	JKL	MN	PRS	TUV	WXY	OQ
1	2	3	4	5	6	7	8	9	0

with Z not used. There is no correspondence between 26 letters and 10 digits. If Z is omitted, there is a correspondence. The set of 25 letters is mapped *into* the set {1, 2, 3, ..., 9, 0} of digits, and *onto* the set {2, 3, ... 9, 0}, the mapping being many-one. Equally, the set of words (in a dictionary or place-name sense) is mapped into a set of sets of digits. The mapping, being many-one, is not useful to the telephone company. The object of their exercise is to get a subset of words such that the first three letters of the words have a one-one mapping into the set of digit triples. For example: EASt→327 and EALing→325; both these words can go into the required subset. But DARtford→327 also; this word must be eliminated in defining the subset for the one-one mapping into digit triples.

7.4. Isomorphism. A main concern of modern mathematics is with systems which may have different labels but which cannot be distinguished in their relevant properties. In the words of Poincaré (1854–1912): 'mathematics is the art of calling different things by the same name'. The point can be made with a simple example. Let X be a set of temperatures $x°$ C. and Y the set of corresponding temperatures $y°$ F.; there is a function or mapping: $y=32+\frac{9}{5}x$. From the point of view of order ($<$), the sets X and Y are the 'same': if $x_1<x_2$ for two members of X, then $y_1<y_2$ in Y. However, if we try to add and multiply temperatures, the sets are not algebraically the 'same' at all: if $x_1=2x_2$ it does not follow that $y_1=2y_2$. Hence, X and Y are the 'same' as long as we do no more than order temperatures. They are 'different' if we attempt to get sums or products of temperatures, which is why we do not do this.

The concept of algebraic sameness or similarity is an important one, and technically it goes by the forbidding name of *isomorphism*.[†] In making the concept explicit and formal, we may appear to be doing no more than stress the obvious. But we find that progress is much easier, and more sure, if we are explicit. For sets X and Y in which a binary operation $*$ is specified, the definition is:

Definition: *An* **isomorphism** *is a one-one mapping* $X \underset{f}{\leftrightarrow} Y$ *of a set* X *onto a set* Y *which preserves the operation* $*$: *if* x_1 *has image* y_1 *and* x_2 *image* y_2 *under* f, *then* $x_1 * x_2$ *has image* $y_1 * y_2$.

The condition of preservation of $*$ can also be put: if $x_1 \leftrightarrow y_1$ and $x_2 \leftrightarrow y_2$, then $(x_1 * x_2) \leftrightarrow (y_1 * y_2)$. More shortly, for the mapping $y = f(x)$:

$$f(x_1 * x_2) = f(x_1) * f(x_2)$$

each side being $y_1 * y_2$. The operation $*$ is usually $+$ or $\times$. For sums: If $x_1 \leftrightarrow y_1$ and $x_2 \leftrightarrow y_2$, then $(x_1 + x_2) \leftrightarrow (y_1 + y_2)$; or

$$f(x_1 + x_2) = f(x_1) + f(x_2).$$

The condition is similar for products.

The term isomorphism refers to the mapping of X onto Y. We can then say that Y is the isomorphic image of X or (simply) that X and Y are *isomorphic* sets. Speaking loosely, we imply that isomorphic sets are substantially the same, indistinguishable from the point of view of the operation in question (though perhaps not for others). The sets are denoted differently and have different interpretations, but they behave algebraically in the same way. They can be described as 'equivalent up to an isomorphism' and, to indicate this, we can write $X \cong Y$.

The first example shows that sets can be isomorphic with respect to two operations at the same time:

(i) $\{1, 2, 3, \ldots\} \cong \{\frac{1}{1}, \frac{2}{1}, \frac{3}{1}, \ldots\}$, both for sums $(+)$ and for products $(\times)$. One set is J^+, the natural numbers or positive integers. The other set is a subset of the set R of all rationals. That they are isomorphic is a consequence of the definition of rationals from integers (2.6). The isomorphism is the justification for writing the rational $\frac{3}{1}$ = the integer 3, and for saying that the integers are included in the rationals: $J \subset R$. From the definition (4.6), the set of finite cardinal numbers is also isomorphic with the set J^+ of positive integers. Hence,

[†] 'Isomorphism' is derived from the Greek: *isos* = equal, and *morphe* = form.

for purposes of sums and products, the positive integer n, the rational number n and the cardinal number n are interchangeable.

The next example illustrates the concept for a relation (that of order), as well as for an operation such as + or × ; it also shows that we must always be careful to see that there *is* an isomorphism:

(ii) $X=\{1, 2, 3, \ldots\}$ and $Y=\{2, 4, 6, \ldots\}$. A one-one mapping from X onto Y is given by $y=2x$ (x a positive integer). The question is: what operations or properties does the mapping preserve. It does preserve both the property of order ($<$) and the operation of summation ($+$). Suppose $n_1 \leftrightarrow 2n_1$ and $n_2 \leftrightarrow 2n_2$. Then $n_1 < n_2$ implies $2n_1 < 2n_2$. Further, $n_1+n_2 \leftrightarrow 2(n_1+n_2)=2n_1+2n_2$ is implied. The mapping does *not* preserve the operation of multiplication ($\times$), for: if $n_1 \leftrightarrow 2n_1$ and $n_2 \leftrightarrow 2n_2$, then $n_1 n_2 \leftrightarrow 2n_1 n_2 \neq 2n_1 \times 2n_2$. For instance: the image of 3 is 6, the image of 4 is 8 and the image of 7 is 14 (adding both sides), but the image of $3 \times 4 = 12$ is $24 \neq 6 \times 8$. Hence $X \cong Y$ for order and for sums, but not for products.

The following example shows that an isomorphism can exist within a single set X, i.e. a one-one mapping of X onto itself, preserving an operation:*

(iii) In the set C of complex numbers, the element $x+iy$ can be put into one-one correspondence with its conjugate $x-iy$, thus defining a one-one mapping of C onto itself. The mapping preserves addition:

$$\begin{aligned}\text{Image of } &\{(x_1+iy_1)+(x_2+iy_2)\}\\ &=\text{Image of } \{(x_1+x_2)+i(y_1+y_2)\}\\ &=(x_1+x_2)-i(y_1+y_2)\\ &=(x_1-iy_1)+(x_2-iy_2)\\ &=\text{Image of } (x_1+iy_1)+\text{Image of } (x_2+iy_2).\end{aligned}$$

Similarly, it can be shown to preserve multiplication. Hence, both for + and for ×, the set of complex numbers is isomorphic with the set of their conjugates.

Two further examples illustrate that an isomorphism can be with respect to one operation in the first set and a different operation in the second set, and that the sets need not have numbers as elements:

(iv) The set $J=\{n \mid n \text{ an integer}\}$ under + is isomorphic with the set $S=\{2^n \mid n \text{ an integer}\}$ under ×. The isomorphism is the one-one mapping $n \leftrightarrow 2^n$, i.e. in $J \underset{f}{\leftrightarrow} S$ the image of n in J is $f(n)=2^n$ in S. If

* It is then called an *automorphism*.

$n_1 \leftrightarrow 2^{n_1}$ and $n_2 \leftrightarrow 2^{n_2}$, then $n_1 + n_2 \leftrightarrow 2^{n_1+n_2} = 2^{n_1} \times 2^{n_2}$. So *addition* of n's in J corresponds in the mapping to *multiplication* of 2^n's in S. The isomorphic sets can be spelled out to show corresponding elements:

$$\{\ldots -3,\ -2,\ -1, 0, 1, 2, 3, \ldots\} \cong \{\ldots \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, \ldots\}$$

where, for example, -3 corresponds to $2^{-3} = \dfrac{1}{2^3} = \dfrac{1}{8}$.

The set J of integers under $+$ is isomorphic, not only with the set of integral powers of 2, but generally with the set of integral powers of any real number under $\times$ (7.9 Ex. 15). J under $+$ is also isomorphic with other kinds of sets under $\times$, e.g. certain sets of transformations with $\times$ taken as repeated application (7.9 Ex. 16).

(v) $\{(x+iy) \mid x \text{ and } y \text{ real}\} \cong \{P \mid P \text{ a point in a plane}\}$, for the operation $+$ between complex numbers and the operation of resultant of vectors OP in a plane. This isomorphism, between the set of complex numbers and the set of all points in a plane, follows from the development of 2.5. It is the justification for showing complex numbers or ordered pairs (x, y) as points on an Argand Diagram.

The most important applications of isomorphisms are to groups and the group operation (usually $+$ or $\times$). A *group isomorphism* is a one-one mapping of a group G onto a set $\bar{G}$, preserving the operation $*$ of G. It can be established that $\bar{G}$ is also a group and that the isomorphism carries over both the identity element and inverses from G to $\bar{G}$. The proof is:

Write $G = \{a, b, c, \ldots\}$ and $\bar{G} = \{\bar{a}, \bar{b}, \bar{c}, \ldots\}$ where $\bar{a}$ is the image of a, $\bar{b}$ of b, … . By the isomorphism: $\bar{a} * \bar{b}$ is the image of $a * b$. $\bar{G}$ is closed under $*$ since G is. Since $a * (b * c) = (a * b) * c$ in G, it follows that their images $\bar{a} * (\bar{b} * \bar{c})$ and $(\bar{a} * \bar{b}) * \bar{c}$ are equal in $\bar{G}$ (associative). The identity e of G has an image $\bar{e}$ in $\bar{G}$; since $e * a = a$, then $\bar{e} * \bar{a} = \bar{a}$ for the images, i.e. $\bar{e}$ is the identity of $\bar{G}$. Finally, the inverse a^{-1} of G has an image $\overline{(a^{-1})}$ in G; since $a^{-1} * a = e$, the same is true of images: $\overline{(a^{-1})} * \bar{a} = \bar{e}$, i.e. $\overline{(a^{-1})} = \bar{a}^{-1}$ the inverse of $\bar{a}$ in G. Hence G is a group with the properties mentioned. So:

THEOREM: *A group isomorphism carries a group G into another* *group $\bar{G}$ as image, carries the identity of G into the identity of $\bar{G}$ and carries an inverse in G into the corresponding inverse in $\bar{G}$.*

It is in this way that one group can be related to, or created from, another group.

Isomorphic groups (or fields) are indistinguishable, from the algebraic point of view of the operations concerned. Some of the examples above have group isomorphisms; two other examples follow:

(vi) The cyclic group $\{1, \omega, \omega^2\}$ of the cube roots of unity is isomorphic with the group of rotations of the equilateral triangle. See 6.4 (ii) above. The group operator is multiplication (×) of complex numbers in one case, successive rotations in the other. Each of them is also isomorphic with the group $\{0, 1, 2\}$ (mod 3) under the group operation of addition (+). The mapping here is similar to that of example (iv) above. The images in $\{1, \omega, \omega^2\} \cong \{0, 1, 2\}$ are: $1 \leftrightarrow 0$, $\omega \leftrightarrow 1$, $\omega^2 \leftrightarrow 2$. Since $\omega^0 = 1$, the exponents of powers on the left are numbers on the right. Then any product of elements on the left is the image of the corresponding sum on the right; e.g.

$$\omega \times \omega^2 = \omega^3 = 1 \quad \text{is the image of} \quad 1 + 2 = 3 = 0$$

using $\omega^3 = 1$ and modulo 3 respectively.

+	Even	Odd
Even	Even	Odd
Odd	Odd	Even

+	0	1
0	0	1
1	1	0

×	Even	Odd
Even	Even	Even
Odd	Even	Odd

×	0	1
0	0	0
1	0	1

(vii) $\{0, 1\}$ (mod 2) $\cong$ {Even, Odd} both for addition (+) and for multiplication (×). See 6.5, example (vi). The result follows from the addition and multiplication tables shown, together with a specification of the mapping:

$$0 \leftrightarrow \text{Even}; \quad 1 \leftrightarrow \text{Odd}$$

which preserves both + and × . This is a double group isomorphism.

Indeed, it can be called a *field isomorphism*, i.e. a one-one mapping preserving both + and ×, carrying over both identities (0 and 1) and both inverses (negatives and reciprocals). In this simplest kind of field, of two elements, the zero elements (0, Even) correspond, and so do the unity elements (1, Odd). Each is its own inverse.

×	1	−1
1	1	−1
−1	−1	1

Notice that the group $\{1, -1\}$ under *multiplication* is isomorphic with the group $\{0, 1\}$ (mod 2) under *addition*, and similarly with the group {Even, Odd} under addition, as shown by the multiplication table for $\{1, -1\}$. The mapping is:

$$1 \leftrightarrow 0 \text{ (Even)}; \quad -1 \leftrightarrow 1 \text{ (Odd)}.$$

The set $\{1, -1\}$ is not a group under addition and group isomorphism does not arise.

7.5. Linear transformations. In 6.3 and 6.4, we considered transformations, with the emphasis on the effect of *different* transformations, e.g. $r(A)$, $s(A)$, $t(A)$, ..., on the *same* object A. A set of such transformations often has the properties of a group. We now consider a *given* transformation T, sending an object A into another object $t(A)$, with attention directed to its effect on *different* objects $A, B, C, \ldots$.

Changing the notation, we write the *transformation* T from one set X into another set Y: $X \underset{T}{\rightarrow} Y$. This means that, if $x \in X$, then there is a unique *image* $t(x) \in Y$ under the transformation T.

A transformation is simply another name for a function or mapping. Algebraically, it is $y = t(x)$; y is a function of x. In diagrammatic terms, it is the mapping T of the set X into the set Y. In simple cases, the functional formulation is to be preferred to the alternative and equivalent expression as a transformation. For example, if X and Y are the same (each the set of positive integers), then the transformation $y = 2x$ is such that each integer x is transformed into the even integer $2x$; the mapping is of points on one line into points on another line (7.3). Here, however, the function $y = 2x$ over the domain of positive integers is the most useful concept.

The opposite may well be true when X and Y are sets of more complicated elements. Let X be the set of ordered pairs of real numbers (x_1, x_2) and Y another set of pairs (y_1, y_2). The transformation: $y_1 = 2x_1$ and $y_2 = x_2$ is again quite simple, corresponding to a magnification of figures (6.3, example (i) above). It is a mapping of points $P(x_1, x_2)$ in the plane Ox_1x_2 into points $Q(y_1, y_2)$ in the plane Oy_1y_2. Alternatively, it maps $P(x_1, x_2)$ into points $Q(2x_1, x_2)$ in the same plane Ox_1x_2. As a function, the transformation is expressed: the pair (y_1, y_2) is a function of the pair (x_1, x_2); this is not particularly convenient.*

More generally, a transformation may map a set of points $P(x_1, x_2, \dots x_n)$ in n dimensions into a set of points $Q(y_1, y_2, \dots y_m)$ in m dimensions. In algebraic terms, the transformation can be shown as giving *each* of $y_1, y_2, \dots y_m$ in terms of the x's $(x_1, x_2, \dots x_n)$. This is usually more helpful than the functional form: $y = f(x)$ where x stands for the n-tuple $(x_1, x_2, \dots x_n)$ and y for the m-tuple $(y_1, y_2, \dots y_m)$. The function $f(x)$ is a function of an n-tuple or vector. It does *not* apply to the components separately; it does *not* mean that $y_1 = f(x_1), y_2 = f(x_2), \dots$.

This concept of a transformation is developed here in the simple case of a linear transformation in two dimensions. The set X of pairs (x_1, x_2) is transformed into the set Y of pairs (y_1, y_2):

$$y_1 = a_{11}x_1 + a_{12}x_2 \quad \text{and} \quad y_2 = a_{21}x_1 + a_{22}x_2 \quad \dots\dots\dots\dots\dots(1)$$

where the x's and y's are real numbers and where the a's are real constants. One example is the simple magnification case of 6.3, i.e. $y_1 = ax_1$ and $y_2 = x_2$ for a constant a. More generally, (1) sends points (x_1, x_2) in the plane Ox_1x_2 into points (y_1, y_2) in the plane Oy_1y_2 in such a way that figures may be contracted, expanded and distorted in various directions. The only certain fixed point under the transformation is the origin $O(x_1 = 0, x_2 = 0)$ which remains unchanged as $O(y_1 = 0, y_2 = 0)$ in (1). Two examples illustrate:

(i) $y_1 = x_1 + x_2$ and $y_2 = \frac{1}{2}x_2$.

The square $ABCD$ in Ox_1x_2 has the following points as vertices, and the transformation sends them into the points $A'B'C'D'$ in Oy_1y_2:

$$A(0, 1) \to A'(1, \tfrac{1}{2}) \qquad C(2, 1) \to C'(3, \tfrac{1}{2})$$
$$B(1, 0) \to B'(1, 0) \qquad D(1, 2) \to D'(3, 1).$$

* But it is useful in one particular case, when the number pairs are complex numbers: $y_1 + iy_2$ as a function of $x_1 + ix_2$. Functions of a complex variable are examined in 7.6.

The shape of the square is changed (into a parallelogram) under the transformation (Fig. 7.5*a*). The transformation can be reversed, to give *ABCD* from *A′B′C′D′*:

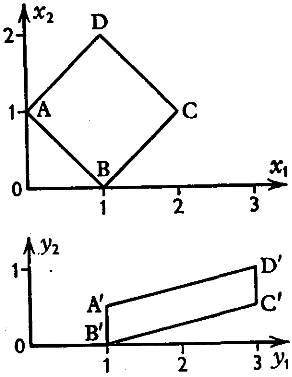


FIG. 7.5*a*

$$x_1 = y_1 - 2y_2 \quad \text{and} \quad x_2 = 2y_2.$$

The transformation (or mapping) is one-one and it has an inverse.

(ii) $y_1 = x_1 + x_2$ and $y_2 = \frac{1}{2}x_1 + \frac{1}{2}x_2$.

Start with the same square *ABCD* in Ox_1x_2. Then:

$$\begin{tikzcd}[row sep=tiny] A(0,1) \arrow[dr] & \\ & A', B'(1, \tfrac{1}{2}) \\ B(1,0) \arrow[ur] & \\ C(2,1) \arrow[dr] & \\ & C', D'(3, \tfrac{3}{2}) \\ D(1,2) \arrow[ur] & \end{tikzcd}$$

The parallelogram *A′B′C′D′* in Oy_1y_2 collapses to two points (Fig. 7.5*b*). Indeed, all points in Ox_1x_2 map into points on the line $y_2 = \frac{1}{2}y_1$ in Oy_1y_2. The transformation cannot be reversed; given y_1 and y_2, the corresponding values of x_1 and x_2 cannot be written. The transformation (or mapping) is many-one, without an inverse.

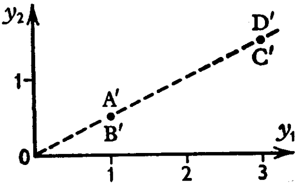


FIG. 7.5*b*

This leads us to look for the inverse of (1). On solving (Appendix A.4):

$$x_1 = \frac{a_{22}y_1 - a_{12}y_2}{a_{11}a_{22} - a_{12}a_{21}}$$

and

$$x_2 = \frac{a_{11}y_2 - a_{21}y_1}{a_{11}a_{22} - a_{12}a_{21}}.$$

This is a linear transformation from (y_1, y_2) to (x_1, x_2), *provided* that:

$$a_{11}a_{22} - a_{12}a_{21} \neq 0 \quad \text{..............................(2)}$$

Hence the transformation (1) is one-one and has an inverse, provided that the constant coefficients in (1) satisfy the condition (2). This is so in example (i). In example (ii), the expression of (2) becomes zero; the condition is not satisfied.

Consider all the different transformations (1), subject to the conditions (2), i.e. take the set of pairs of equations as the *a*'s are given all possible real numerical values which satisfy (2). The set can easily be checked to have the properties of a group under the operation ×

(successive applications of transformations from the set). The critical property is the existence of an inverse for each transformation of the set, assured by the condition (2). Linear transformations (1) which satisfy (2) are called *non-singular*. The result is that non-singular linear transformations, with coefficients from the field of real numbers, form a group under multiplication. It is an example of what is called a *full linear group*.

7.6. Conformal transformations. If the number pair (y_1, y_2) is a function of the number pair (x_1, x_2), then as a transformation each of y_1 and y_2 is given in terms of x_1 and x_2. The linear case is (1) of 7.5 above. The expressions for y_1 and y_2 are, in general, specified separately and independently; there is no necessary connection between them. There are, however, particular cases where the two expressions are linked. One way of linking them is by combining x_1 and x_2 into a complex number (x_1+ix_2) and by transforming into another complex number (y_1+iy_2). The transformation then appears as a *function of a complex variable*. Further, since the two expressions in the transformation are linked, it is called a *conformal transformation*. The function can still be of any form but, to provide illustrations, only two simple cases of polynomial form are considered here.

The notation is changed so that the complex number from which we start is written $z=x+iy$, where x and y are real numbers. The transformation is specified by $Z=f(z)$ where $Z=X+iY$ is the transformed version of $z=x+iy$. On separating the elements of the number pairs, X is expressed in terms of x and y, and Y in terms of x and y; this is the process of 'equating real and imaginary parts'. The two expressions are linked through the form (here a polynomial) adopted for f.

The simplest case is the linear function of a complex variable: $Z=az+b$ where a and b are (real) constants. Hence:

$$X+iY=a(x+iy)+b=(ax+b)+i(ay)$$

and the conformal transformation is

$$X=ax+b \quad \text{and} \quad Y=ay \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

If the transformation (1) is regarded as a mapping of points (x, y) in the plane Oxy into points (X, Y) in the plane OXY, and if $a>0$, then the mapping is a magnification in the direction Ox and the *same* magnification in the direction Oy, together with a simple shift (by b) in the Ox direction. The conformal nature of the transformation appears in the equal magnifications used. For example:

$$Z=\tfrac{3}{2}z-\tfrac{1}{2}$$

or $$X=\tfrac{1}{2}(3x-1) \quad \text{and} \quad Y=\tfrac{3}{2}y$$

corresponds to a 50 per cent magnification in both directions and a horizontal shift to the left (by $\tfrac{1}{2}$). The square $ABCD$ is blown up to the square $A'B'C'D'$ as shown in Fig. 7.6*a*.

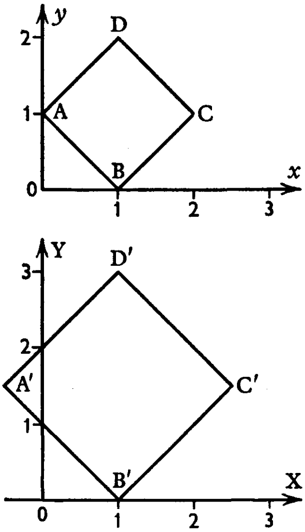


Fig. 7.6*a*

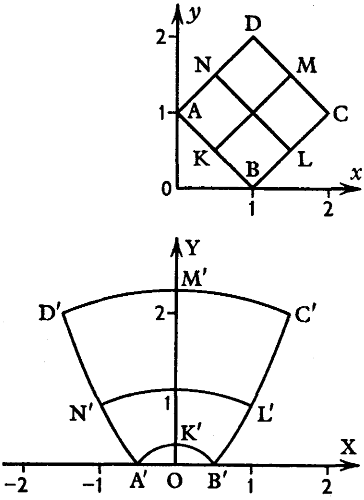


Fig. 7.6*b*

A simple non-linear function of a complex variable is: $Z=\tfrac{1}{2}z^2$

i.e. $$X+iY=\tfrac{1}{2}(x+iy)^2=\tfrac{1}{2}(x^2-y^2)+ixy$$

or $$X=\tfrac{1}{2}(x^2-y^2) \quad \text{and} \quad Y=xy \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

As a mapping of (x, y) in Oxy into (X, Y) in OXY, this transformation distorts a figure composed of straight lines into a curvilinear figure. The conformal transformation (1) is linear and preserves straight lines; the conformal transformation (2) is non-linear and lines are sent into curves. Fig. 7.6*b* illustrates by showing the transformation of the square $ABCD$ into $A'B'C'D'$. To assist in the identification, the mid-points $KLMN$ are also shown, transformed into $K'L'M'N'$. For example, CMD are collinear:

$$C(2, 1) \quad M(3/2, 3/2) \quad D(1, 2).$$

To get C', put $x=2$, $y=1$ in (2) and obtain:

$$X=(2^2-1^2)/2=3/2 \quad \text{and} \quad Y=2\times 1=2.$$

Hence, on working out for all three points:

$$C'(3/2, 2) \quad M'(0, 9/4) \quad D'(-3/2, 2)$$

which are not collinear. The conformal nature of the transformation (2) appears in the preservation of a certain symmetry in the figures.

7.7. Order. In 6.7 ordering is taken first as a primitive concept, i.e. $a<b$ means a precedes b. For a field, in which differences are defined, order becomes more precise; it is a property of positiveness: $a<b$ means $(b-a)$ positive. The concept of a relation now provides a more general notion of order.

A preliminary comment on the notation for order is appropriate. Though we do, in fact, define an ordering by reference to a relation denoted generally by R, we have always in mind that the order is according to $<$ (less than) or $\leqslant$ (less than or equal to). There are alternative and equivalent notations in use in either case. It is a matter of choice whether we write $a<b$ (a less than b) or $b>a$ (b greater than a); both notations mean precisely the same thing. As long as we know what we are doing, it is a great convenience in practice to be able to switch between $a<b$ and $b>a$. In the same way, we can interchange $a\leqslant b$ and $b\geqslant a$ at will. In developing basic concepts, when there are quite enough real distinctions to keep in mind, we must avoid this duplication of notation. Here we confine ourselves to $a<b$ or $a\leqslant b$. Any relation R of an ordering is to be interpreted *either* as '$<$' *or* as '$\leqslant$'.

$S=\{a, b, c, \ldots\}$ may be ordered in one or other of three forms:

(i) *Complete Ordering*. There are two possibilities only:

$$\textit{either } a<b \quad \textit{or} \quad b<a.$$

(ii) *Partial Ordering*. There are three possibilities:

$$a<b \quad \textit{or} \quad a=b \quad \textit{or} \quad b<a.$$

(iii) *Weak Ordering*. There are four possibilities:

$$a<b \quad \textit{or} \quad a=b \quad \textit{or} \quad b<a \quad \textit{or} \quad a \text{ and } b \text{ not comparable.}$$

Write R for the relation of the ordering. R is *complete* if it holds, one way or the other, between every pair of elements of S: either aRb or bRa. Not all orderings have a complete relation, e.g. R is not complete in a weak ordering. The *negation* R' of R is the relation between a and b which holds when a and b are not related aRb, i.e. $aR'b$ means $\sim(aRb)$.* Consider each ordering in turn:

(i) In a complete ordering, take R as $<$ so that aRb means $a<b$. Then $aR'b$ means $a \nless b$, i.e. $b<a$, the only alternative. The relation R is complete, since either aRb $(a<b)$ or bRa $(b<a)$ holds. Equally, R' is complete, since either $aR'b$ $(b<a)$ or $bR'a$ $(a<b)$ holds.

(ii) In a partial ordering, take R as $\leqslant$ so that aRb means $a \leqslant b$. Then $aR'b$ means $a \nleqslant b$, i.e. $b<a$, the only remaining possibility. Again R is complete, since either $aRb\,(a \leqslant b)$ or $bRa\,(b \leqslant a)$. But R' is *not* complete, since $aR'b\,(b<a)$ and $bR'a\,(a<b)$ do not exhaust the possibilities ($a=b$ not allowed for).

(iii) In a weak ordering, take R as $\leqslant$ again. Then $aR'b$ is also an alternative: $a \nleqslant b$ means *either* $b<a$ *or* a and b not comparable. Neither R nor R' is complete.

The concept of equivalence (7.2) is relevant to ordering. In a complete ordering, no elements of S are related by equivalence. In a partial or weak ordering, there are equivalent elements in S so that S can be partitioned into equivalence classes. Two elements a and b of S may belong to the same equivalence class, in which case $a=b$. Where the two orderings differ is in the properties of elements not in the same equivalence class. For a partial ordering, the equivalence classes can themselves be ordered completely. Hence, if a and b are not in an equivalence class, then either $a<b$ or $b<a$. This is not so for a weak ordering. The equivalence classes cannot be put in a complete order; there is always the possibility that a and b are not comparable. See 7.9 Ex. 26.

7.8. Properties of order. The three properties of equivalence (7.2) are to be examined in the wider context of ordering. One is the *transitive* property: if aRb and bRc, then aRc. This is assumed to hold throughout.†

* In terms of sets, the relation R is a subset of all ordered pairs (a, b) of the Cartesian product $S \,.\, S$, e.g. those for which $a \leqslant b$. Then R' is the complementary subset, all pairs not related by R, e.g. those for which $b<a$.

† Though not considered here, non-transitive relations are of considerable interest. They may even describe what can be called ordering.

Another is the *reflexive* property: aRa logically true, i.e. if a and b are identical, then aRb. For example, the relation $\leqslant$ is reflexive. The opposite property is *irreflexive*: aRa logically false, i.e. if aRb, then a and b cannot be identical. For example, the relation $<$ is irreflexive.

The remaining property is concerned with any *symmetry* which exists between aRb and bRa. This can best be regarded as a consequence of the reflexive and transitive properties assumed for R. There are two cases to consider. First, suppose that R is reflexive and transitive. Then aRb and bRa may both hold, i.e. equivalence of a and b $(a=b)$ is possible. The typical case arises when R is $\leqslant$, which is reflexive $(a\leqslant a)$ and transitive. It is possible that $a\leqslant b$ and $b\leqslant a$, meaning $a=b$. The set S has a subset which is an equivalence class. Second, suppose that R is irreflexive and transitive. Then aRb and bRa cannot both hold. For, if they do, then aRa by the transitive property, and this is ruled out by the irreflexive property. The typical case is when R is $<$, which is irreflexive ($a<a$ false) and transitive. It is not possible that $a<b$ and $b<a$ both hold. This is the anti-symmetric property.

We can now write formal *definitions*, and obtain *properties*, for an ordered set $S=\{a, b, c, \ldots\}$ of elements of any kind:

(i) *Complete Ordering*. The set S is completely ordered by R if R is an irreflexive, transitive and complete relation. Hence R is anti-symmetric. The relation R can then be written $<$. Hence, for any a, b and c in S:

Property	*In terms of R:*	*With R written <:*
Irreflexive	aRa logically false	$a \nless a$
Anti-symmetric	if aRb, then $\sim(bRa)$	if $a<b$, then $b\nless a$
Transitive	if aRb and bRc, then aRc	if $a<b$ and $b<c$, then $a<c$
Complete	if $\sim(aRb)$, then bRa	if $a\nless b$, then $b<a$

The negation R' has precisely the same properties. This is because $aR'b$ means $\sim(aRb)$ or bRa. With R written $<$, aRb is $a<b$ and $aR'b$ is $b<a$; these are the only possibilities.

(ii) *Partial Ordering*. The set S is partially ordered by R if R is a reflexive, transitive and complete relation. Hence symmetry is possible, i.e. aRb and bRa may both hold (a and b equivalent). The relation R can be written $\leqslant$:

Property	*In terms of R:*	*With R written $\leqslant$:*
Reflexive	aRa logically true	$a \leqslant a$
Equivalence	if aRb and bRa, then a and b are equivalent	if $a \leqslant b$ and $b \leqslant a$, then $a = b$
Transitive	if aRb and bRc, then aRc	if $a \leqslant b$ and $b \leqslant c$, then $a \leqslant c$
Complete	if $\sim(aRb)$, then bRa	if $a \nleqslant b$, then $b \leqslant a$

The negation R' is irreflexive, anti-symmetric and transitive, but *not* complete. Since $aR'b$ means $\sim(aRb)$, then, with R written $\leqslant$, $aR'b$ means $b<a$. The reason why R' is not complete is that $aR'b$ ($b<a$) and $bR'a$ ($a<b$) do not exhaust the possibilities; they omit the case $a=b$.

(iii) *Weak Ordering.* The set S is weakly ordered by R if R is reflexive and transitive. Again symmetry is possible, i.e. aRb and bRa may both hold (a and b equivalent). The relation R can be written $\leqslant$:

Property	*In terms of R:*	*With R written $\leqslant$:*
Reflexive	aRa logically true	$a \leqslant a$
Equivalence	if aRb and bRa, then a and b are equivalent	if $a \leqslant b$ and $b \leqslant a$, then $a = b$
Transitive	if aRb and bRc, then aRc	if $a \leqslant b$ and $b \leqslant c$, then $a \leqslant c$
Incomplete	if $\sim(aRb)$, then bRa may or may not hold	if $a \nleqslant b$, then $b \leqslant a$ or a and b are not comparable

The negation R' is a relation with alternatives: $aR'b$ means $b<a$ or a and b not comparable. R' is irreflexive, anti-symmetric and transitive. Neither R nor R' is complete; there is always the possibility a and b not comparable to confuse R and the possibility $a=b$ to confuse R'.

The ordering of a set $S=\{a, b, c, \ldots\}$ bears directly on the question whether S can be *scaled* or not, i.e. on the question of the 'ordinal measurement' of the elements of S. The scaling of S is achieved if S is completely or partially ordered; no scaling is possible if S is weakly ordered. A completely ordered S has no equivalence classes (of more than one element) and there is no *indifference relation* between the elements of S in the scaling. A partially ordered S has equivalence classes so that, in scaling S, certain elements must be counted as indifferent on the scale. A typical case is the set of rational numbers which can be completely scaled if duplication of rationals is eliminated but which is scaled with indifference relations if duplication is not eliminated. For example, the rationals $\frac{3}{2}$, $\frac{6}{4}$, $\frac{9}{6}$, ... form an indifference subset in the scaling of the rationals.

The concept of isomorphism applies to ordered sets. The sets $X=\{x_1, x_2, x_3, \ldots\}$ and $Y=\{y_1, y_2, y_3, \ldots\}$ are isomorphic or *similarly ordered* if there is a one-one mapping $x_1 \leftrightarrow y_1$, $x_2 \leftrightarrow y_2$, $x_3 \leftrightarrow y_3$, ... which preserves ordering by the relation R:

$$\text{if } x_1 R x_2, \quad \text{then} \quad y_1 R y_2.$$

For a complete ordering, $x_1 < x_2$ implies $y_1 < y_2$; for a partial ordering, $x_1 \leqslant x_2$ implies $y_1 \leqslant y_2$. The same scaling can be applied to similarly ordered sets.

7.9. Exercises

1. Take $X=\{1, 2\}$ and $Y=\{1, 2, 3, 4, 5\}$ and show that the relation

$$R=\{(x, y) \mid x \in X, \quad y \in Y, \quad y=x^2\}$$

consists of 2 out of the 10 elements of $X \,.\, Y$. Replace X by {Even, Odd} and show that $X \,.\, Y$ has 10 elements with 5 in the relation R' given by the statement $yR'x$: 'y is x'. Show that the domain of R and of R' is the whole of X.

2. If X and Y are each the set of all real numbers, explain why the statements $y=|x|-1$ and $y=\surd(x^2)-1$ give the same relation R. Represent R graphically.

3. Illustrate that a variable can be related to a discrete number by considering 'y is the least integer not less than x' for $X=\{x \mid x \text{ a real number, } 0<x\leqslant 4\}$ and $Y=\{1, 2, 3, 4\}$. Show that X is the domain and Y the range of the relation.

4. *Relations with compound statements.* If X and Y each comprise all real numbers, the relation $R=\{(x, y) \mid x \in X, \ y \in Y, \ x^2+y^2=1, \ x>0\}$ is specified by the conjunction of two statements. Compare with

$$R'=\{(x, y) \mid x \in X, \quad y \in Y, \quad x^2+y^2=1\}$$

where X is all positive, and Y all real numbers. In what sense are R and R' the same? Represent graphically as in Fig. 7.1*b*.

5. In the tribe of 4.1, example (iii), X is the set of 4 upper-class males and Y the set of 3 lower-class males. A relation R is defined by 'y is of the same generation as x'. Show that R comprises 5 out of 12 pairs in $X \,.\, Y$.

6. Show that the relations R of Exs. 1 and 2 both give y as a function of x; and that the relations R' of Exs. 1 and 4 do not give y as a function of x but do give x as a function of y. Show that the correspondence is two-two in the relation of Ex. 5 and that no function is defined.

7. *Step-functions.* The relation 'y is the least integer not less than x' is defined on the domain X of all positive real numbers. Show that y is a function of x with the property that one and the same value of y is given for all x in $0<x\leqslant 1$, another single value of y for all x in $1<x\leqslant 2$, and so on. Show graphically why this can be described as a step-function.

8. Show that 'x and y have the same father' is an equivalent relation in any set of people. Consider the relation 'y is the brother of x', showing that it is symmetric (but not reflexive) in a set of men but that it fails even to be symmetric in a set comprising men and women.

9. It might be argued that the reflexive condition is not needed in the definition of an equivalent relation R: if xRy, then yRx (symmetry); but, if xRy and yRx, then xRx (transitivity); i.e. xRx follows and need not be specified. But this means xRx *if* there is a y such that xRy, *not* xRx for all x. Illustrate by reference to the relation 'x and y are both even' in the set of all integers.

* 10. *Circular relations.* The condition: if zRy and yRx, then xRz, defines a circular relation. Show that R is an equivalence relation if and only if it is both reflexive and circular. Illustrate with 'x and y have the same father'.

11. Prove the converse of the partitioning theorem of 7.2: if S is partitioned into subsets S_x, then an equivalence relation yRx is defined in S. To prove, take yRx as 'x and y belong to the same subset S_x' and establish that R satisfies the conditions for an equivalence relation.

12. Show that the statement $y=2x$ gives a mapping of $X=\{1, 2, 3, \ldots\}$ *into* itself but a mapping of $X=\{x \mid x \text{ a real number}\}$ *onto* itself. In illustrating the importance of specifying the domain of a function (mapping), indicate that the difference here is due to the fact that X is a field in the second (continuous) case, but not in the first (discrete) case.

* 13. Partition the set J of integers (a group under $+$) into equivalence classes J_r ($r=0, 1, 2, \ldots n-1$) by the relation 'x and y have the same remainder on division by n'. Show that the set of canonical forms is the set of integers (mod n). The J_r's are residue classes (3.9 Ex. 18). Show also that J_0 is a subgroup of J but not the others. The J_r's are cosets in the group J (6.9 Ex. 12).

* 14. Continuing, show that the set of residue classes J_r is isomorphic with the set of integers (mod n), preserving both $+$ and $\times$, and deduce that the J_r's make up a field if n is prime. For this, define:

$$J_r+J_s=\text{set of sums of an element of } J_r \text{ and an element of } J_s$$

and similarly for $J_r\times J_s$. Then show that $J_r+J_s=J_{(r+s)}$ and $J_rJ_s=J_{(rs)}$ where $(r+s)$ and (rs) are sum and product respectively of r and s (mod n).

15. Show that $\{n \mid n \text{ an integer}\}$ under $+$ is isomorphic with $\{\alpha^n \mid n \text{ an integer}\}$ under $\times$, for any real α. Deduce that 10^x under $\times$ behaves like x under $+$, for x an integer (and more generally, see Chapter 12). This is the basis of logarithms.

16. *Group of translations.* Consider the group $n(A)$ of translations, i.e. a shift of n to the right, where n is an integer (6.4). Show that this group under $\times$ is isomorphic with the group J of integers under $+$.

17. Show that $\{0, 1, 2, 3, 4\}$ (mod 5)$\cong\{0, 2, 4, 6, 8\}$ (mod 10) under $+$ but not under $\times$.

18. *Cyclic groups.* Show that $\{a, a^2, a^3, \dots a^n\}$ where $a^n = 1$ (identity) under $\times$ is isomorphic with the set of integers (mod n) under $+$.

19. *Automorphism.* If a and b are any rationals, show that

$$\{a + b\surd 2\} \cong \{a - b\surd 2\},$$

preserving both sums and products. This is an automorphism, a one-one mapping of the field $R(\surd 2)$ onto itself.

* 20. *Another derivation of complex numbers.* Show that the field of polynomials, mod $x^2 + 1$, is isomorphic with the set $a + bx$, where $x^2 = -1$, preserving both $+$ and $\times$. (Here the coefficients of the polynomials and the pair (a, b) are to be taken as real numbers.) See 3.9 Ex. 19 and 20. Deduce that the field of complex numbers can be *defined* from the integral domain of polynomials $f(x)$ over the field of real numbers, by taking remainders on division of $f(x)$ by $x^2 + 1$ and by interpreting x as i $(i^2 = -1)$.

* 21. *Homomorphism.* Modify the definition of isomorphism (7.4): a *homomorphism* is a *many-one* mapping $X \xrightarrow[F]{} Y$ preserving the operation $*$, $F(x_1 * x_2) = F(x_1) * F(x_2)$. An isomorphism is then a particular case of the more general mapping of a homomorphism. Illustrate by showing that there is a homomorphism between the group J of integers under $+$ and the cyclic group $\{1, i, -1, -i\}$ under $\times$, the many-one mapping being $n \to i^n$. Contrast the set $\{i^n \mid n \text{ an integer}\}$ with the set $\{\alpha^n \mid n \text{ an integer}\}$ which is isomorphic with J for α real (Ex. 15).

22. Show that $y_1 = \frac{3}{2}x_1$ and $y_2 = x_1 - x_2$ has inverse $x_1 = \frac{2}{3}y_1$ and $x_2 = \frac{2}{3}y_1 - y_2$, and that it sends the square of Fig. 7.5a into a parallelogram.

23. Compare the transformation $y_1 = a_{11}x_1 + a_{12}x_2 + b_1$, $y_2 = a_{21}x_1 + a_{22}x_2 + b_2$ with (1) of 7.5, showing that it has the same effect on a figure in the plane Ox_1x_2, except that it sends O into $O'(b_1, b_2)$ in Oy_1y_2, i.e. it includes a shift or change of origin. If the transformation is non-singular $(a_{11}a_{22} - a_{12}a_{21} \neq 0)$ show that the inverse exists:

$$x_1 = \frac{a_{22}(y_1 - b_1) - a_{12}(y_2 - b_2)}{a_{11}a_{22} - a_{12}a_{21}}, \quad x_2 = \frac{a_{11}(y_2 - b_2) - a_{21}(y_1 - b_1)}{a_{11}a_{22} - a_{12}a_{21}}.$$

24. *Affine group.* Consider the set of non-singular transformations of Ex. 23, for all real a's and b's $(a_{11}a_{22} - a_{12}a_{21} \neq 0)$. The subset for which $b_1 = b_2 = 0$ is a group (the full linear group of 7.5). Show that the subset for which $a_{11} = a_{22} = 1$ and $a_{12} = a_{21} = 0$ is also a group, the group of translations (Ex. 16). Then show that the complete set is a group which is not commutative. See 6.3, example (i). The complete group is called an affine group.

25. The function $Z = \dfrac{1}{z}$ of a complex variable gives a conformal transformation, mapping (x, y) into (X, Y) by $X = x/(x^2 + y^2)$ and $Y = -y/(x^2 + y^2)$. Trace the transformation of the figure $ABCD$ of Fig. 7.6b in this case.

26. *Weak ordering.* Order a set $S=\{a, b, c, \ldots\}$ according to the relation aRb: 'I like b at least as much as a'. Show that there are four possibilities for a pair a and b: I like a and b equally (indifference, aRb and bRa both true), *or* I like b better (b preferred, aRb true but not bRa), *or* I like a better (a preferred, bRa true but not aRb), *or* I cannot choose (non-comparable, neither aRb nor bRa true). Deduce that R gives a weak ordering of S.

27. The relation R is reflexive in the set $S=\{a, b, c, \ldots\}$. Show that

$$(aRb \wedge bRa)$$

is a relation $a\bar{R}b$ between a and b and that $\bar{R}$ is an equivalence relation in S. Deduce that $\bar{R}$ serves to partition a partially ordered set S into equivalence classes, i.e. separating into classes the equal $(a=b)$ members of S. See the second property of partial ordering in 7.8.

CHAPTER 8

GEOMETRIES

8.1. Various geometries. The properties of figures in space are the concern of geometry and trigonometry.* We deal here mainly with 'plane geometry' in two dimensions. More generally, we need to visualise figures in three dimensions and to imagine them in more than three dimensions. A vast amount of material has accumulated in geometry and, without referring to more than a small fraction, we can attempt to provide a logical basis for it all.

A first distinction is on the *method of treatment* in geometry. This can be *synthetic* in the sense of logical deduction from geometric axioms on the lines of the famous system of Euclid (*circa* 300 B.C.). It can be *analytic* in the sense that points are associated with number pairs, with geometric properties translated into algebraic relations. There is then a unification of geometry and algebra or analysis; which swallows up the other is a matter for argument.

Another distinction is concerned with the *content* of geometry, with the kind of properties to be established. Elementary geometry deals with the metric aspect of figures; it is designed to apply to the everyday world of distance, angle and area. The idea here is that the figures of metric geometry are unchanged in shape by certain transformations which may be described as 'rigid motions', i.e. translations, rotations and reflections. It is not necessary that the figures should actually be 'moved'. A reflection, for example, is the transformation corresponding to looking at a figure in a mirror. Moreover, translations and rotations can be expressed in terms of two or more reflections (8.9 Ex. 1). Hence in metric geometry, figures are *invariant* under certain *transformations*. Again geometry is linked to algebra, and in particular to groups of transformations. A vast extension of

* 'Geometry' is derived from the Greek: *ge* = earth, and *metria* = measuring; similarly, for measurement of triangles, 'trigonometry' comes from: *tri* = three, *gonia* = angle, and *metria* = measuring.

geometry is made by following up the question: what properties of figures are invariant under this or that group of transformations? In this way, we are led to such subjects as *projective geometry* (8.7 below).

A third distinction gets down to fundamentals, the *axiomatic basis* of geometry. We may think that, in metric geometry, we are talking about actual space, about actual points and lines. This is not so. We postulate abstract concepts of points and lines and develop abstract properties of these abstractions. In applications to everyday life, we must remember that points and lines on paper (and still more on the surface of the earth) are approximate realisations of abstract points and lines. The question is: what axiomatic basis is appropriate? The idea that there is only one set of axioms and only one geometry has been in discard since Einstein developed his famous theory of relativity. It can no longer be said that Euclid's axioms are correct and all others wrong, or even a waste of time. We can safely consider all kinds of axioms, and hence various geometries, Euclidean and non-Euclidean, both as interesting academic exercises and with the thought that they may even have practical uses.

Is the axiomatic basis of geometry to be given in geometric or in algebraic terms? Until late in the nineteenth century, all respectable geometers followed a purely geometric formulation for metric and projective geometry alike, often going far out of their way to do so. The algebraic or analytic treatment was played down very severely; it was useful for illustration and (sometimes) for getting at otherwise difficult proofs. More recently, a better balance has been struck between the purely geometric and the algebraic treatments. There is indeed a good case to be made out for an algebraic basis of all geometries, for geometry to be absorbed into algebra.

We have already used the link between algebra and geometry, admittedly only for illustrative and graphical purposes. There is an isomorphism between a set of number pairs (x, y) and a set of complex numbers $(x+iy)$, as a matter of straight algebra. However, either can be associated (made isomorphic) with a set of points P, or with a set of vectors OP, in a plane. We need to get the geometric aspect here onto a firmer basis. There is clearly some risk of confusion, with four different sets isomorphic one with another. We find that the vector is the appropriate concept to take as a basis, to be related to a point P and a number pair (x, y). Complex numbers and their representation on an Argand Diagram are best left on one side, as a useful but subsidiary application.

8.2. Metric geometry and vectors. We take for granted the main results of mensuration in elementary geometry and trigonometry (Appendix A.7 and A.8).

One feature of metric geometry is that metric properties are *invariant* under the group of transformations of 'rigid motions', i.e. translations, rotations and reflections. If one of these transformations is applied to the vertices of a triangle ABC, sending A into A', B into B' and C into C', then the triangle $A'B'C'$ is congruent with the triangle ABC.

Another feature is that there is a very considerable amount of *duality* between points and lines.* The *incidence* of a point P and a line p means *both* that P lies on p *and* that p goes through P. Further, given two points P and Q, a unique line r joins P and Q, defining a unique segment or length PQ along r. The dual result is: given two lines p and q, a unique point R is obtained as the intersection of p and q, defining a unique angle (p, q) between p and q. A line is the dual of a point and an angle is the dual of a segment. A triangle is specified either by three points ABC or (as the dual) by three lines abc.

There is one troublesome exception to this duality. Two points P and Q always define a line r. On the other hand, two lines p and q define a point R, *except* where p and q are parallel. It is clearly worth while to eliminate this nuisance; this will be done later. There is also a difficulty in definition, shown up in the duality between points and lines. In defining the line r joining two given points P and Q (and the length of the segment PQ), how do we know that we go from P to Q to get the length PQ? We might start from P, go the wrong way and never get to Q. Or, to put the matter another way, what points on r are 'between' P and Q? Now look at the same difficulty in dual form. In defining the point R of intersection of two given lines p and q, we say we have a unique angle (p, q) between the lines. But which angle? There are four of them and (even with opposite angles rated as equal) two different ones. Or, how do we know which lines through R are

* Whenever there is no risk of ambiguity, we use 'line' for 'straight line'.

'between' p and q to make the angle (p, q)? Fig. 8.2*a* illustrates. Clearly this concept of 'betweenness' is implicit in metric space, just as 'order' is in the field of real numbers.

The way out of the difficulty is to introduce the idea of a *vector*, a concept uniting points and lines and embracing both length and direction (angle). A vector has both given length and given direction but with no fixed position; it refers to the relative position of two points. In short, a vector is the *relative and directed distance* between two points in space.* It is unchanged if the points are shifted (translated) in the same way in space.

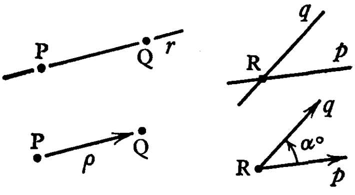


FIG. 8.2*a*

Hence, the vector PQ is the segment of line from P to Q, having both length (ρ) and direction (from P to Q). Given P and Q, we now know which way we are going, which points are between P and Q. So we get the length ρ from P to Q in the vector PQ. The vector QP is different: the same length ρ but in the opposite direction from Q to P. We write $QP = -PQ$, one vector the negative of the other. Similarly given two lines p and q intersecting in R, specify them as vectors from P with given directions. Then we have a specified (unique) angle $\alpha°$ for the angle (p, q), that measured anti-clockwise from the (directed) p to the (directed) q. Fig. 8.2*a* illustrates.

In the end, we find that vectors are basic for defining 'space' and that they are a strictly algebraic concept. We can build the geometry of space on a firm algebraic foundation. To do this, we first look for the properties we wish vectors to have, and we then use them, suitable abstracted, as the axiomatic basis for defining vector space. The properties (illustrated in Fig. 8.2*b*) are:

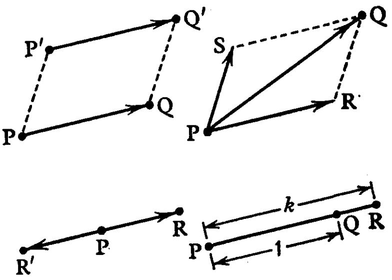


FIG. 8.2*b*

* 'Vector' is a Latin word: *vector*, *-oris* = carrier, from *veho*, *vexi*, *vectum* = carry. Hence the original idea of a vector is a displacement, carrying one point to another, by a certain distance in a certain direction.

(i) *Equivalence*: PQ and $P'Q'$ are equivalent (equal) if they have the same length and direction. So opposite sides of a parallelogram give equivalent vectors. This expresses the idea of a vector as a directed distance but with no fixed position.

(ii) *Addition:* if PQ is the diagonal of the parallelogram $PRQS$ formed from the vectors PR and PS, then PQ is the sum of PR and PS:

$$PQ = PR + PS.$$

The sum is the *resultant* of the vectors, an idea arising in mechanics, e.g. for the sum or resultant of forces. It is also the way in which complex numbers are summed on an Argand Diagram. The corresponding process of subtraction follows; the sum $PQ = PR + PS$ gives two differences:

$$PQ - PR = PS \quad \text{and} \quad PQ - PS = PR.$$

The *zero* vector is PP, of no length. The *negative* $(-PR)$ is the vector PR' such that $PR + PR' = PP$ and P is the mid-point of RR'. (If the parallelogram $PRQS$ has $PR = PS$ and if the angle between PR and PS opens out to 180°, then the resultant PQ collapses on P.) Hence the negative $(-PR)$ is a vector of the same length as PR but reversed in direction. It is easily checked (8.9 Ex. 4) that the difference:

$$PS = PQ - PR = PQ + (-PR)$$

and similarly that $PR = PQ - PS = PQ + (-PS)$.

(iii) *Multiplication by a scalar*: if k is a real number (scalar), then $PR = kPQ$ is a vector obtained as follows. If $k > 0$, PR is in the same direction as PQ but of k times the length, i.e. a contraction of $PQ (k < 1)$, PQ itself $(k = 1)$, or an expansion of $PQ (k > 1)$ as in Fig. 8.2*b*. If $k = 0$, the vector of zero length is obtained. If $k < 0$, write $k = -\kappa$ where $\kappa > 0$. Then κPQ is a contraction or expansion of PQ. So $PR = kPQ = -(\kappa PQ)$ is the same contraction or expansion, but in the opposite direction. In particular, if $k = -1$, $PR = -PQ$ is the vector PQ reversed in direction. Hence, multiplication by k contracts or expands a vector, in the same direction $(k > 0)$ or with direction reversed $(k < 0)$.

All this is in part familiar and in part strange. The familiar aspect is that vectors form an *additive group*, a commutative group under the operation of addition here defined (8.9 Ex. 5). It is commutative since PQ is either $PR + PS$ or $PS + PR$ from the parallelogram. The new aspect is the operation of *scalar multiplication* for expansion or contraction. This is *not* multiplication of vectors; we have said nothing, and need say nothing, about the product of two vectors.*

8.3. Vector spaces. We have now the idea of a vector in two dimensions, something with length and direction. In generalising, we abandon the limitation to two dimensions and, at first, we leave over the features of length and direction as needing more precise definition. These features may be thought of as 'purely geometric' concepts but, in fact, they are as yet no more than visual impressions. We find, perhaps to our surprise, that geometric 'space' in a very general sense can be specified in terms of vectors, without reference to length and direction. Moreover, the definition is entirely algebraic, an extension of the concept of a group.

We start with a set V of elements $u, v, w, \ldots$ which we propose to call *vectors*. The vectors of V are *entities* of various kinds, not at all necessarily numbers. They may be single numbers and quite often they are number pairs (or n-tuples); but they can be things without reference to numbers at all. The essential requirement is that an operation of addition $(+)$ is defined for vectors so that V is an additive group, i.e. V has all the properties (including the commutative one) of a group under addition. Addition can be of the nature of a 'resultant' of two vectors but we leave it open. Being an additive group, V includes a *zero vector* which we denote by 0 as usual.

An additive group was the starting point in defining a field (6.5). In that case, we created a system of double composition by adding a second group operation (multiplication). We now proceed on a different tack, still aiming to make V a system of double composition. We supplement the additive group V by a second operation, but this

* In general there is no question of writing the product of two vectors as another vector; a vector space is not a field with an operation × as well as +. There is an exception: a vector written as a number pair (x, y) in two dimensions can be interpreted as a complex number $(x + iy)$, represented on an Argand Diagram and subject to multiplication by the rule of 2.5. In short, a *vector space in two dimensions* can be identified as a *field of complex numbers* when we wish to multiply the vectors. There is no such identification in three or more dimensions. This is why we said in 8.1 that complex numbers are best left on one side in dealing with vectors; they apply only to the special case of two dimensions.

is now a different one: scalar multiplication. V will not be a field like real numbers but something basically different. The reason for the difference is that, for scalar product, we need to lay our hands on an outside set of elements, $F = \{a, b, c, \dots\}$, which we propose to call *scalars*. F is both a distinct set and a set of different nature from V. We require F to be a field under its own operations of + and ×, and typically F is a field of numbers. Here we take F as the field of real numbers. Scalars are then real numbers, taken from the field F of all real numbers, and quite outside the set V of vectors.

Hence the second operation in V is scalar multiplication: given an element u of V, select a scalar a from F and form the scalar product au as another vector. So one vector u of V gives another vector au of V on multiplication by a scalar a from F. With an eye on the two-dimensional representation of 8.2, we specify the properties we require of scalar products. There are four of them:

The Operational Rules of Scalar Multiplication

For the set $V = \{u, v, w, \dots\}$ of vectors multiplied by scalars from

$$F = \{a, b, c, \dots\}.$$

Rule	*Scalar products*
S1. Closure	au belongs to V
S2. Unit scalar	$1u = u$
S3. Associative	$a(bu) = (ab)u$
S4. Distributive	$a(u+v) = au + av$ and $(a+b)u = au + bu$

We start with closure: a vector times a scalar always gives a vector (S1). Then the particular function of the unit scalar 1 is laid down: multiplication by 1 leaves a vector unchanged (S2). The associative rule (S3) is that, in multiplying by two scalars, it doesn't matter whether we multiply in two stages or whether we first multiply the scalars and then apply the product direct to the vector. The distributive rule (S4) has two parts; one shows the distribution of one scalar over the sum of two vectors, and the other the distribution of the sum of two scalars over one vector.

These rules are so framed that they form an 'economical' list of properties. They can be specified as the axioms for scalar products. They are complete, consistent and independent of each other. In this matter, two observations are appropriate. First, the commutative rule does not arise here; it is not that it fails to hold but rather that we do not need it. We are operating in V and, in doing so, we dip outside for a scalar a to multiply u to give au. We are *not* operating in F; and so, given a in F, we never dip outside for u to give ua in F. Secondly, the zero scalar 0 in F is related to the zero vector 0 in V: $0u=0$. Here, we write the same symbol 0 for two different zero elements, one in F and one in V. This simplifies matters and the context makes it clear always which zero is which. Further, the scalar -1 in F is related to negative vectors in V: $(-1)u=-u$. As operational rules both $0u=0$ and $(-1)u=-u$ may be added to S2 above. They can, however, be derived from the other rules (8.8 Ex. 6 and 8).

Pulling together the threads of the discussion, we define:

DEFINITION: *The set* $V=\{u, v, w, \ldots\}$ *is a* **vector space** *over the field* $F=\{a, b, c, \ldots\}$ *of* **scalars,** *if the two operations of addition of vectors and of multiplication of a vector by a scalar are defined in V so that V is a commutative group under addition and so that scalar products satisfy the properties* S1, S2, S3 *and* S4 *above.*

A field is a specialised set, satisfying a long list of operational rules for sums and products; typical cases are the rational, real or complex numbers. A vector space is also a specialised set, satisfying a long list of operational rules for sums and scalar products. The two lists in part overlap (for sums) but they are also different (products and scalar products differ in their nature).

As a matter of attaching a label, we now say that the general concept of 'space' in geometry is any set of vectors V with these specialised properties. We define 'vector space' algebraically, and *call* it 'geometric space'. We no longer have the kind of set typically represented by numbers. But we do have a link with numbers since the set V of vectors is defined over the field F of numbers. It is this link which enables us to use numbers, in the form of co-ordinates, in geometry.

Vector space is a very abstract concept. The question remains: how do we distinguish one space from another, with particular reference to their possible applications? The answer turns on what additional properties we allocate to vector space, and in particular what concept of 'distance' in space we choose to define. Various Euclidean and non-Euclidean spaces are obtained by varying the concept of distance used in the general vector space. We have seen to it that our abstract vector space has all the required properties of the two-dimensional vectors of 8.2, except for length and direction. These latter are still to be specified.

8.4. Euclidean space. A particular case of Euclidean space, that appropriate to the two dimensions of the plane, is here developed. Once this is done, the extension to more than two dimensions is easily achieved. Euclidean space is the general vector space V of 8.3 with two specific properties added. One is that each vector of V is represented by an ordered n-tuple of real numbers, drawn from the same field F as the scalars. Here, with $n=2$ (two dimensions), take a vector as an ordered pair (x, y) of real numbers. Specify the operations of addition and scalar multiplication:

DEFINITION: *The product of (x, y) by the scalar k is (kx, ky) and the sum of*

$$(x_1, y_1) \quad \textit{and} \quad (x_2, y_2) \quad \textit{is} \quad (x_1+x_2, y_1+y_2) \quad \text{...........(1)}$$

These are obvious interpretations of the two operations. It is easily checked that the vectors form an additive group and satisfy the rules S1–S4 of 8.3. The additive group has a zero, the vector (0, 0).

The other property is the definition of distance and angle:

DEFINITION: *If (x_1, y_1) and (x_2, y_2) are two vectors of the space V, the* **distance** *d and* **angle** *$\theta°$ between them are given by the scalars:*

$$d=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2} \quad \textit{and} \quad \cos\theta=\frac{x_1x_2+y_1y_2}{\sqrt{x_1^2+y_1^2}\sqrt{x_2^2+y_2^2}} \quad \text{.........(2)}$$

Note that d is the positive square root of a positive scalar. In giving the angle $\theta°$, the scalar shown measures an angle in much the same way as a thermometer reading measures a temperature. With various values of the scalar are associated corresponding measures of angle in the unit called 'degrees'. Two specific 'scaling points' are needed: the zero angle 0° for which the scalar is 1, and the right-angle 90° for which the scalar is 0. The scalar is to be identified as the trigonometric ratio (cosine) and it is so written in (2); later we must make sure that the notation cos θ is justified.

There is no reference yet to any application to physical space, e.g. the plane of this paper or the surface of the earth. We have to attach appropriate labels and to check that the abstract properties correspond with physical properties (suitably idealised). The vector (x, y) can be called the *point* P and the whole set of vectors V, as various selections of real numbers x and y are made, makes up the *plane* of points P. The zero vector $(0, 0)$ is labelled the *origin* O, the point in the plane from which distances can be conveniently measured. Alternatively, the (algebraic) vector (x, y) can be called the *geometric vector* OP from O to P, in line with the idea of a vector discussed in 8.2.

The rules (1) for scalar products and sums can be re-interpreted. If P is (x, y), write $Q(kx, ky)$ as the product of P by the positive scalar k. By (2), the angle between P and Q is 0° and the distance OQ is k times the distance OP. This is what we mean when we say that Q is on the line through O and P. Q is between O and $P\,(k<1)$, at $P\,(k=1)$ or beyond $P\,(k>1)$; the last is illustrated in Fig. 8.4*a*. For negative k, Q still lies on the line OP but on the opposite side of O from P; the distance OQ is $|\,k\,|$ times the distance OP.

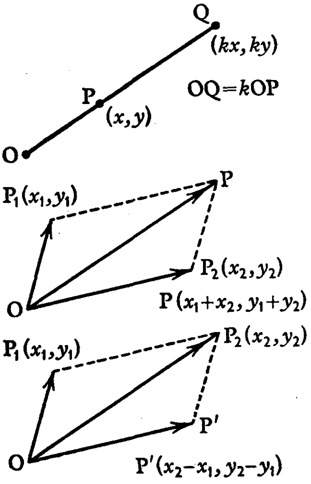


FIG. 8.4*a*

Let P_1 be (x_1, y_1) and $P_2(x_2, y_2)$. The sum is the point $P(x, y)$ such that $x=x_1+x_2$ and $y=y_1+y_2$. As a definition, say that OP_1PP_2 is a parallelogram, the sum OP being the diagonal, obtained from the sides OP_1 and OP_2. The (geometric) vector sum $OP=OP_1+OP_2$ is the resultant of the separate (geometric) vectors OP_1 and OP_2. Fig. 8.4*a* illustrates. Similarly, the difference $P'(x', y')$ between P_1 and P_2 has $x'=x_2-x_1$ and $y'=y_2-y_1$. Then the (geometric) vector difference $OP'=OP_2-OP_1$ means that OP_2 is the resultant of OP_1 and OP'.

So far, all geometric vectors are from the fixed point O to any point P. We can agree, however, to call the vector from P_1 to P_2 as the same as the difference vector $OP'=OP_2-OP_1$. They are opposite sides of a parallelogram. We can then use the familiar *triangle of vectors* (Fig. 8.4*b*). We have:

$$OP_2 = OP_1 + OP' = OP_1 + P_1P_2 \quad \text{and} \quad P_1P_2 = OP' = OP_2 - OP_1.$$

We can read vectors around the triangle of Fig. 8.4*b*, e.g. OP_2 is the sum of OP_1 and P_1P_2, or P_1P_2 is the difference OP_2 less OP_1.

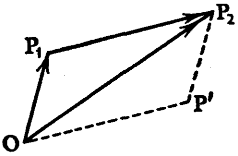


FIG. 8.4*b*

We are now ready to map the algebraic vectors (x, y) onto a set of points P in an abstract plane, with familiar applications to a physical plane. The origin O is inserted as a starting point, and then two other points, A (1, 0) and B (0, 1), to serve as measuring rods for distances. By (2), each of A and B is unit distance from O. By (2), also, the angle between A and B (or between the geometric vectors OA and OB) is $\alpha°$ where $\cos\alpha = 0$. The angle is 90°; OA and OB are perpendicular. The conventional siting of A and B is to take OA horizontal (A unit distance to the right of O) and OB vertical (B unit distance above O) as shown in Fig. 8.4*c*. The set of points $M(x, 0)$ for real x defines a directed line or *axis* Ox passing through O and A. For, by (1), $(x, 0)$ is the scalar x times $(1, 0)$; by (2), the distance OM is x and the angle between A and M is 0°. The points M $(x, 0)$ make up an ordered set on the directed line Ox, to the right of $O\,(x>0)$ or to the left $(x<0)$, increasing in distance from O as $|x|$ increases. Similarly, the set of points N $(0, y)$ for real y defines an axis Oy through B. The axes are perpendicular.

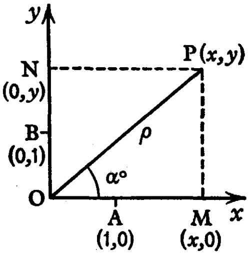


FIG. 8.4*c*

For points P (x, y), neither x nor y zero, consider first the case of positive values $(x>0, y>0)$ to get what we know as the *positive quadrant* of the plane. If M is $(x, 0)$ and N $(0, y)$, then OM and ON sum to OP where P is (x, y), i.e. OP is the diagonal of the parallelogram formed by OM and ON. But OM and ON are at right angles and the parallelogram is what we know as a rectangle. Hence to locate P (x, y): take M a distance x along Ox, N a distance y along Oy, complete the rectangle and get P as the corner opposite O. P is located by means of 'co-ordinates' (x, y) and with reference to axes Ox and Oy.

Further details elaborate the picture. The geometric vector OP can be said to have *length* ρ, where ρ is the distance from O (0, 0) to P (x, y). Then $\rho=\sqrt{x^2+y^2}$ by (2). The vector also has *direction*, given by the angle $\alpha°$ which OP makes with Ox. Define trigonometric ratios from the triangle OPM:

$$\left.\begin{aligned}\cos\alpha &= \frac{OM}{OP}=\frac{x}{\rho}=\frac{x}{\sqrt{x^2+y^2}}\\ \sin\alpha &= \frac{MP}{OP}=\frac{y}{\rho}=\frac{y}{\sqrt{x^2+y^2}}\end{aligned}\right\}\quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

where OM, OP and MP are distances and where MP is

$$\sqrt{\{(x-x)^2+(y-0)^2\}}=y$$

by (2). But the angle $\alpha°$ is that between OP and OA, i.e. between vectors (x, y) and $(1, 0)$, so that (2) gives $\cos\alpha=x/\sqrt{(x^2+y^2)}$. We have agreement; the trigonometric ratios tie in with the basic property (2) of angles between vectors. The notation (2) is justified.

The results (3) can be expressed most conveniently:

$$x=\rho\cos\alpha \quad\text{and}\quad y=\rho\sin\alpha \quad\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

giving the 'co-ordinates' x and y in terms of the length ρ of the vector OP and the angle $\alpha°$ it makes with Ox (Fig. 8.4*c*). P can be located, either by specifying x and y, or by specifying ρ and α. The pair (x, y) are the *Cartesian co-ordinates* of P, named after Descartes (1596–1650), and the pair (ρ, α) are the *polar co-ordinates* of P. They are related by (4).

The extension to the whole set of vectors V and to the whole plane Oxy is a matter of allowing for negative values of x and y in (x, y). See 8.9 Ex. 14–16. This involves the negative of the vector OP (multiplication by the scalar -1) as a vector of the same length but opposite in direction. Trigonometric ratios need to be extended to apply to angles which are not acute in such a way that the relations (4) are preserved (Appendix A.7 and A.9). In the end, the variation of P over the whole plane corresponds to the set of vectors (x, y) for any real x and y. An ordering is involved, i.e. the ordered set of points $(x, 0)$ on the axis Ox and the similar ordered set of points $(0, y)$ on the other axis Oy. The ordering is precisely that of the field of real numbers.

Euclidean space, in the two dimensions of 'plane geometry', is defined algebraically by writing a vector (x, y) to satisfy the requirements of a vector space over the field F of real numbers, and by taking addition, scalar products, distances and angles as defined by (1) and (2) above. The extension to any number of dimensions is immediate. In three dimensions, for application to 'solid geometry', vectors are taken as triples (x, y, z), where x, y and z are real numbers from F. The definitions (1) and (2) only need appropriate extension to allow for three terms instead of two. Points, vectors, distances and angles in three dimensions are referred to three axes Ox, Oy and Oz mutually at right angles. Further, in abstract Euclidean space of n dimensions, vectors are n-tuples $(x_1, x_2, \dots x_n)$ and there are n axes mutually at right angles. The algebraic development is perfectly capable of supporting n dimensions, even when $n > 3$; the difficulty is to visualise the resulting points and vectors.

In general, a vector space of n-tuples $(x_1, x_2, \dots x_n)$ over a field F is denoted by $V_n(F)$. It depends both on the number n of dimensions and on the field, e.g. of real numbers, used for co-ordinates and scalars alike. If appropriate definitions of distance and angle are added, $V_n(F)$ becomes Euclidean space of n dimensions, denoted by $E_n(F)$. The case examined here is $E_2(F)$ over the field F of real numbers.*

8.5. Non-Euclidean spaces. Two properties of Euclidean space are particularly to be noticed. Consider the triangle OP_1P_2 of Fig. 8.4*b* above, where the vector OP_2 is the sum of the vectors OP_1 and P_1P_2. Denote vector lengths:

$$|OP| = \text{length of } OP = \surd(x^2 + y^2).$$

where $P(x, y)$ is the vector. Then, from the vector sum

$$OP_2 = OP_1 + P_1P_2,$$

it follows:

$$|OP_2| \leqslant |OP_1| + |P_1P_2|.$$

* An attempt is sometimes made to generalise Euclidean space, by refraining from specifying that the vectors are n-tuples of scalars from F, and by replacing the definition of distance and angle, (2) above, by certain scalars called 'inner products' with appropriate properties. The Euclidean space so obtained turns out to be isometric (i.e. isomorphic, preserving distance and angle) with the space $E_n(F)$ here obtained for n-tuple vectors. It is essentially the same; nothing of importance is gained.

The equality holds only if OP_1 and OP_2 are in the same direction. From this comes the Euclidean property: a line is the shortest distance between two points.

Further, the vectors OP_1 and OP_2 have a *unique* difference (in the additive group of vectors): the vector $OP_2 - OP_1 = P_1P_2 = OP'$, where P' completes the parallelogram in Fig. 8.4*b*. Hence the *unique* OP' is the parallel through O to the given vector P_1P_2. The result: a unique parallel to a given line can be drawn through an outside point. This is another of Euclid's postulates.

Going back to general vector space, we can throw in a definition of distance which is different from that chosen for Euclidean space and, at the same time, we can define lines so that they are still the shortest distance between two points (vectors). However, the property about parallels through a given point (Euclid's postulate) will no longer be necessarily true with a new definition of distance. Two variants of the property can arise:

(i) There is *no* parallel to a given line through an outside point.

(ii) There are *several* parallels to a given line through an outside point.

The spaces so obtained are non-Euclidean. In case (i), they are called *elliptic*, and often named after Riemann (1826–66) who investigated their properties. In case (ii), they are called *hyperbolic*, and often named jointly after the two mathematicians Bolyai (1802–60) and Lobachevsky (1793–1856)* who first established in detail the construction of non-Euclidean geometries.

Two illustrative examples are given here, without attempting to go into the details of the specification and properties of the spaces. Both examples, however, are of interest in that they can have practical applications.

The first example is that of *spherical geometry*, the geometry of two-dimensional points and lines on the surface of a sphere (in Euclidean space of three dimensions). It is thus possible to translate from non-Euclidean space of two dimensions into Euclidean three-dimensional terms. In spherical geometry, lines are the shortest distances between points on the surface of a sphere, i.e. they are great circles of the sphere. Three lines intersecting in a spherical triangle are shown in Fig. 8.5*a*. For two lines to intersect in a unique

* This is the Lobachevsky rendered into song by Tom Lehrer.

point (and three lines in a unique triangle), we must adopt the convention that a point and its antipodal counterpart (on the other side of the sphere) are the same. One and only one line can be drawn through two distinct points; but for a unique distance between the points, we need the convention that the shorter arc (of the great circle) is taken. Then *every* pair of lines intersects in a point without exception. There are *no* parallel lines. If we draw a circle on the sphere parallel to a great circle like AA' (e.g. by drawing 'tram lines' round the equator), the second circle is *not* a great circle, it is *not* the shortest distance between points on it and, in fact, it is *not* a line. Hence we have an elliptic non-Euclidean geometry, of type (i). One feature of triangles (spherical triangles) is that the sum of the angles is greater than 180°. For example, one line can be the equator (AA'), a second can be a polar great circle (BB'), and the third can be another polar great circle (CC' shifted to intersect BB' at the poles, B and B'). The triangle so formed has *two* right-angled angles (on the equator AA') and a third (non-zero) angle at a pole.

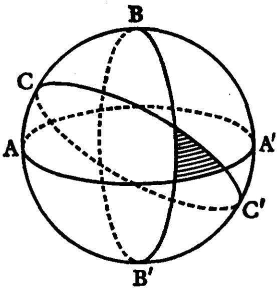


Fig. 8.5*a*

The second example is that of the two-dimensional *geometry of the inside of a circle*. The circle is in Euclidean space of two dimensions; this space extends outside the circle whereas the non-Euclidean space is confined within it. Again a translation into Euclidean terms is possible. A line in Euclidean space, as the shortest distance between points, is the path of a light ray with a constant velocity in a homogeneous medium. Assume that the inside of the circle is a non-homogeneous medium such that the velocity of light varies from point to point. Then a line inside the circle, as the path travelled by a light ray, appears to be curved to the Euclidean outsider. For example, Fig. 8.5*b* illustrates the case where the velocity of light is proportional to the distance of the point from the circumference of the circle; lines are arcs of (Euclidean) circles cutting the boundary at right angles. Suppose BC is a given line and A a point not

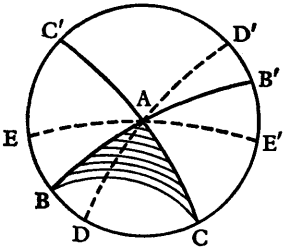


Fig. 8.5*b*

on BC; BB' and CC' are lines through B and C respectively, which meet at A, making a triangle ABC. Another line through A (such as DD') in the shaded angle BAC cuts BC. But a line through A (such as EE') in the other angle does *not* cut BC, i.e. it is parallel to BC, and there are many such parallels. We have a hyperbolic non-Euclidean geometry, of type (ii). In this geometry, the sum of the angles of a triangle is less than 180°. The triangle ABC shown has zero angles at B and C and a third angle of approximately 90°. This is the non-Euclidean geometry of Poincaré (1852–1912); it has relevance to certain theories of the actual universe of the astronomer and physicist.

8.6. Co-ordinate geometry. In 8.5 we have taken a line as the shortest distance between two points. We now abandon this idea. Instead, in returning to Euclidean space, we exhibit a line as a special case of a 'locus', the geometric equivalent of the algebraic concept of a 'relation'.

From sets X and Y, each consisting of all real numbers, form the Cartesian product $X \,.\, Y = \{(x, y) \mid x \in X, y \in Y\}$ as the set of all ordered pairs (x, y) of real numbers. A *relation* is a proper subset of $X \,.\, Y$, specified by some statement yRx, and denoted by

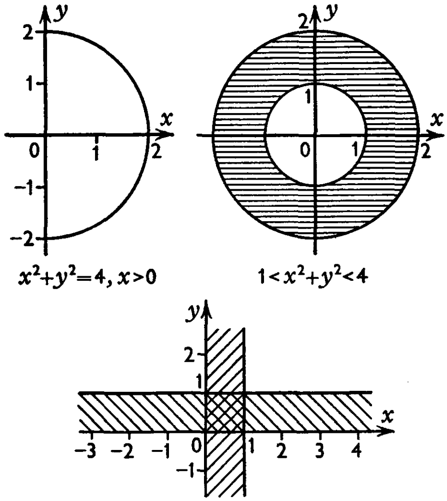
$$\{(x, y) \mid x \in X,\ y \in Y,\ yRx\},$$

or more shortly by yRx. A *locus* is a set of points (x, y) in the plane Oxy, the points for which some statement yRx holds. If an actual plotting of the locus is made on a physical plane, the result is a *graph*; it can be regarded either as the graph of the relation or as the graph of the locus. Hence, given a relation yRx, the geometric representation is a locus and an actual plotting is a graph. Co-ordinate geometry is the study of properties of loci by means of an algebraic treatment of corresponding relations.

The relation/locus concept is a very wide one; it is just any subset of real number pairs (x, y) or of points in a plane. As shown in 7.1, it may be specified by a single equation or inequality, e.g. $x^2 + y^2 = 1$ is the locus of points on a circle, and $x^2 + y^2 < 1$ is the locus of points within a circle. It is quite possible, indeed common, to specify a relation/locus by two or more equations or inequalities. The locus is then the intersection or union of two or more subsets, one for each of the equations/inequalities. In such cases, we often speak of the locus as the 'solution set' obtained by the intersection (or union) of the separate sets (see 6.8 above). Consider three examples:

(i) $x^2+y^2=4$ and $x>0$.

The locus of this relation consists of all points on a semi-circle, centre at O and radius 2, as shown in Fig. 8.6*a*. It is the intersection of the set of points on the circle ($x^2+y^2=4$) and the set of points in the half-plane to the right of Oy ($x>0$).



$x^2+y^2=4$, $x>0$

$1<x^2+y^2<4$

$0<x<1$ *or* $0<y<1$

FIG. 8.6*a*

(ii) $1<x^2+y^2<4$.

The locus is the set of points in the ring-like space between two circles (both with centre at O, of radius 1 and 2 respectively). It is the intersection of the set of points within one circle ($x^2+y^2<4$) and the set of points outside the other circle ($x^2+y^2>1$).

(iii) $0<x<1$ *or* $0<y<1$.

This is a relation which specifies 'or' rather than 'and'. The locus corresponds to the union (rather than the intersection) of two sets. One set is the vertical 'band' of points between Oy and the line parallel to Oy and distance 1 from it ($0<x<1$), shown with one hatching in the graph (Fig. 8.6*a*). The other set is a similar horizontal 'band' of points shown with another hatching in the graph ($0<y<1$). The relation specifies that one or other (or both) of the statements ($0<x<1$, $0<y<1$) holds. The locus is the union of the two sets, i.e. all points hatched in the diagram. Compare: $0<x<1$ *and* $0<y<1$, a relation and locus which are the intersection of the two sets, as cross-hatched in the graph.

A *function* $y=f(x)$ is a particular kind of relation, that for which a unique y corresponds to any given x in its domain. The locus is then a set of points in Oxy with the property that vertical lines (parallel to Oy) cut it in no more than a single point. None of the loci of the above examples is of this kind; the relations are not functions. If a function is given algebraically, then we can write y explicitly in terms of x in some way, as illustrated:

(iv) $x^2+y^2=4$ and $y<0$.

Here $y=-\sqrt{(4-x^2)}$ is the explicit function. The locus is a semi-circle below Ox (centre at O and radius 2) as graphed in Fig. 8.6*b*.

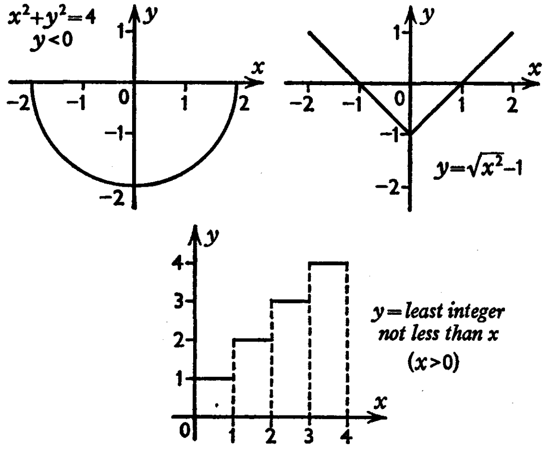


FIG. 8.6*b*

(v) $y=\sqrt{(x^2)}-1$.

This is in explicit form and it can also be written $y=|x|-1$, where $|x|$ is the absolute value of x. The locus consists of points on two half-lines meeting at $(0, -1)$, as graphed.

(vi) $y=1$ when $0<x\leqslant 1$; $y=2$ when $1<x\leqslant 2$; $y=3$ when $2<x\leqslant 3$; ... .

The explicit statement of the function is:

$$y=\text{least integer not less than } x \quad (x>0).$$

It is a *step-function*. The locus is a set of points which ascend in steps, as graphed; at each step, the point of the locus is on the lower rung ($x=1, 2, 3, \ldots$).

The simplest general relation is the *linear relation* specified by $ax+by+c=0$, where a, b and c are real *parameters*, given in the form of two ratios $a:b:c$, and such that a and b are not both zero. If $a\neq 0$, the two parameters are b/a and c/a. If $b\neq 0$, they can be written a/b and c/b. The corresponding locus is a line:

DEFINITION: *A (straight)* **line** *is the locus of points (x, y) in the plane Oxy satisfying the linear relation $ax+by+c=0$, where a, b and c are given real numbers (a and b not both zero).*

Distinguish first a special case, and then the more general case:

Case $b=0$. Hence $a\neq 0$ and $ax+c=0$ or $x=(-c/a)$ i.e. x is defined for only one real number $(-c/a)$ and y is then any real number. The locus is a line parallel to Oy, as graphed in Fig. 8.6c in the particular case $x=1$.

Case $b\neq 0$. Hence $ax+by+c=0$ can be written $y=(-a/b)x+(-c/b)$ and y is a function of x defined on the set of all real numbers. The locus is a line not parallel to Oy, as graphed for $x+y-1=0$, or $y=1-x$, in Fig. 8.6c.

The main properties of a line can be summarised:

THEOREM: *The line $ax+by+c=0$ contains the point*

$$P(\lambda x_1+\mu x_2, \lambda y_1+\mu y_2)$$

if it contains the points $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$, for any real numbers λ and μ such that $\lambda+\mu=1$. Conversely, if a locus contains the point P whenever it contains P_1 and P_2, then the locus is a line.

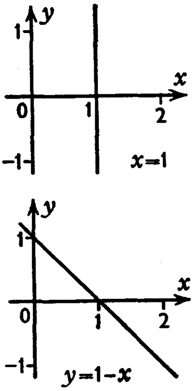


FIG. 8.6c

The proof is simple algebra. For the direct result:

$$ax_1+by_1+c=0 \quad \text{and} \quad ax_2+by_2+c=0$$

since P_1 and P_2 are on the line. Hence:

$$\lambda(ax_1+by_1+c)+\mu(ax_2+by_2+c)=0 \quad \text{for any } \lambda \text{ and } \mu$$

i.e. $$a(\lambda x_1+\mu x_2)+b(\lambda y_1+\mu y_2)+c(\lambda+\mu)=0$$

i.e. $$a(\lambda x_1+\mu x_2)+b(\lambda y_1+\mu y_2)+c=0 \quad \text{since } \lambda+\mu=1.$$

Hence, P is on the line, as required. For the converse:

Fix any two points $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ on the locus (whatever it is). Then $P(x, y)$ is a point on the locus where

$$x = \lambda x_1 + \mu x_2 \quad \text{and} \quad y = \lambda y_1 + \mu y_2 \quad (\lambda + \mu = 1).$$

Let λ and μ be any real numbers (subject only to $\lambda + \mu = 1$) and determine a relation between x and y by eliminating λ and μ:

$$y_2 x = \lambda x_1 y_2 + \mu x_2 y_2 \quad \text{and} \quad x_2 y = \lambda x_2 y_1 + \mu x_2 y_2$$

i.e. $y_2 x - x_2 y = \lambda(x_1 y_2 - x_2 y_1)$ by difference.

Similarly: $y_1 x - x_1 y = -\mu(x_1 y_2 - x_2 y_1)$.

Subtract and put $\lambda + \mu = 1$:

$$(y_2 - y_1)x - (x_2 - x_1)y = (x_1 y_2 - x_2 y_1) \quad \text{.................}(1)$$

Since (1) is a linear relation, the locus is a line. Q.E.D.

The result can be expressed in terms of vectors and their sums and scalar products. Given two vectors $\mathbf{P}_1(x_1, y_1)$ and $\mathbf{P}_2(x_2, y_2)$, write the vectors which are scalar products by λ and μ respectively:

$$\lambda\mathbf{P}_1\text{: } (\lambda x_1, \lambda y_1) \quad \text{and} \quad \mu\mathbf{P}_2\text{: } (\mu x_2, \mu y_2).$$

Add these vectors to give the vector $\mathbf{P}$:

$$\mathbf{P} = \lambda\mathbf{P}_1 + \mu\mathbf{P}_2 \quad \text{i.e. } \mathbf{P}\text{: } (\lambda x_1 + \mu x_2, \lambda y_1 + \mu y_2).$$

The theorem then states: the line through $\mathbf{P}_1$ and $\mathbf{P}_2$ contains $\mathbf{P} = \lambda\mathbf{P}_1 + \mu\mathbf{P}_2$ if and only if $\lambda + \mu = 1$. This is, indeed, an alternative definition of a line (see 8.9 Ex. 13), i.e. a definition in geometric rather than algebraic terms.

A line through $O(0, 0)$ is of particular interest. If the relation is $ax + by + c = 0$, $x = 0$ and $y = 0$ must satisfy it, i.e. $c = 0$. A line through O is the locus $ax + by = 0$ for any given ratio $a : b$. By the theorem in vector form, if $\mathbf{P}_1(x_1, y_1)$ is a point on the line (in addition to O), then $\mathbf{P} = \lambda\mathbf{P}_1 + (1 - \lambda)O = \lambda\mathbf{P}_1$ is on the line for any λ. Hence a line through O is described by scalar multiples $\lambda\mathbf{P}_1$ of a given vector $\mathbf{P}_1$. Scalar multiples differ only in length, not in direction.

The scalars λ and μ determine which points are 'between' P_1 and P_2 on the line through P_1 and P_2, and which are 'outside' P_1 and P_2. Write $\mu = 1 - \lambda$, so that

$$\mathbf{P} = \lambda\mathbf{P}_1 + (1 - \lambda)\mathbf{P}_2 \quad (\text{any } \lambda)$$

gives a point P on the line P_1P_2. The cases are:

$\lambda = 0$: P at P_2; $\lambda = 1$: P at P_1;

$0 < \lambda < 1$: P between P_1 and P_2;

Otherwise: P not between P_1 and P_2.

The property of between-ness corresponds to the ordering of the real numbers. If P_1 is (x_1, y_1) and $P_2(x_2, y_2)$, then P is

$$\{\lambda x_1 + (1-\lambda)x_2,\ \lambda y_1 + (1-\lambda)y_2\}.$$

Hence, in the case of Fig. 8.6*d*, M is between M_1 and M_2 on Ox if $x_1 < \lambda x_1 + (1-\lambda)x_2 < x_2$ i.e. if $0 < \lambda < 1$. Similarly, N is between N_1 and N_2 on Oy if $0 < \lambda < 1$. It is in this sense that P is between P_1 and P_2 if $0 < \lambda < 1$.

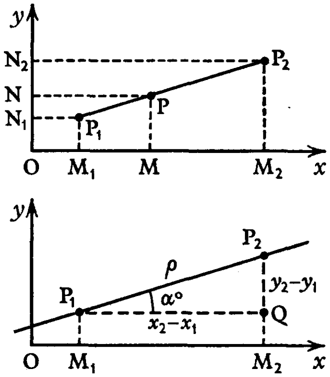


FIG. 8.6*d*

Case $b = 0$, line parallel to Oy, is easily disposed of. If $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ are two points on the line, then $x_1 = x_2$. The general case, $b \neq 0$, has $x_1 \neq x_2$ for any two points P_1 and P_2 on the line. This is now assumed.

The *slope* of the line $ax + by + c = 0$, or $y = (-a/b)x + (-c/b)$, is defined to be the ratio $(-a/b)$. $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ are any two points on the line so that:

$ax_1 + by_1 + c = 0$ and $ax_2 + by_2 + c = 0$.

Subtract: $a(x_2 - x_1) + b(y_2 - y_1) = 0$

i.e. $$\text{slope} = -a/b = (y_2 - y_1)/(x_2 - x_1) \quad \ldots\ldots(2)$$

Further, if the length P_1P_2 is ρ and if the direction of the line is given by the angle $\alpha°$ it makes with Ox, then:

$$\rho = \sqrt{\{(x_1 - x_2)^2 + (y_1 - y_2)^2\}} \quad \text{and} \quad \tan\alpha = \text{slope} = (y_2 - y_1)/(x_2 - x_1).$$

Here $\tan\alpha$ is a trigonometric ratio, equal to $\sin\alpha/\cos\alpha$. Fig. 8.6*d* illustrates and see 8.9 Ex. 3.

To determine the equation of a particular line, we need to be given *either* two points on the line, *or* one point on the line and the slope. From (1) and (2), it follows that the equations are:

$$\left.\begin{array}{ll} \text{Two points } P_1(x_1, y_1) \text{ and } P_2(x_2, y_2)\text{:} & y - y_1 = \dfrac{y_2 - y_1}{x_2 - x_1}(x - x_1) \\ \text{One point } P_1(x_1, y_1) \text{ and slope } m\text{:} & y - y_1 = m(x - x_1) \end{array}\right\} \ldots\ldots(3)$$

Each of (3) is a linear relation $ax + by + c = 0$ with particular forms for the parameters.

A non-linear locus is to be specified by the non-linear relation yRx

which defines it. Some of the loci are 'curves', such as circles, parabolas, ellipses and hyperbolas, as examined in plane geometry. To illustrate simply:

$$(x-a)^2+(y-b)^2=r^2 \quad (a, b, r \text{ given real numbers, } r>0)$$

is a relation with three parameters a, b and r. The corresponding locus is defined to be a *circle*. The familiar geometric property of a circle then follows at once. If P is any point (x, y) on the locus, and if Q is the given point (a, b), then:

$$PQ=\sqrt{\{(x-a)^2+(y-b)^2\}}=r>0 \quad \text{by the relation.}$$

Hence P is a given distance r from the point Q. A circle is the locus of a point which is a given distance r from a given point (a, b). The parameters a, b and r are at choice; the centre of the circle is (a, b) and the radius r. If $a=b=0$ and $r=2$, then $x^2+y^2=4$ is a circle centre O and radius 2.

8.7. Projective geometry. Distances and angles, defined in Euclidean space, are invariant under a particular group of transformations, i.e. translations, rotations and reflections. These transformations are rigid motions not affecting the shape of any figure or of any locus as a set of points. In particular, the distance $\sqrt{\{(x_1-x_2)^2+(y_1-y_2)^2\}}$ between two points $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ is invariant (8.9 Ex. 23). Other kinds of transformations do not leave distances or angles invariant; they distort a 'figure'. Here we consider briefly one such group of transformations (projections) to illustrate the fact that invariant properties depend on the particular type of transformation applied.

In projective geometry, we view the properties of points in two dimensions from a vantage point in three dimensions. The idea is rather similar to that of using complex numbers (and an Argand diagram in two dimensions) to get properties of real numbers (points on a line in one dimension). We have extra freedom in dealing with real numbers; for example, the perfect square x^2+y^2 is exhibited as the product of conjugate complex numbers $(x+iy)(x-iy)$, and the negative $(-x)$ is obtained from x by two operations of multiplication by i (rotation through 90°).

Consider points such as P and figures such as the triangle ABC on a plane Π. In three-dimensional space, select another plane Π' and some point Q not on either Π or Π' as in Fig. 8.7*a*. Join Q to the point P on Π and let the line QP cut Π' in the point P'. Then P on Π is said to have *image* P' on Π' by *projection* from the given centre Q. This is a transformation of points P on Π into corresponding images P' on Π'. Applying the transformation to the triangle ABC on Π, we get a triangle $A'B'C'$ on Π'. Projection from Q sends points into points and lines into lines. But the shape of a figure on Π is altered when projected into its image on Π'. Distances and angles are not invariant, so that (for example) the triangle ABC and its image $A'B'C'$ can be of quite different shapes.

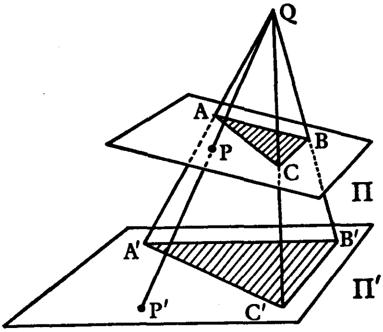


FIG. 8.7*a*

Given the centre Q but a whole set of planes Π, Π', Π'', ..., we obtain a set of transformations (projections) from points on any one plane to images on another plane. The product of two transformations can be defined as successive projections, from points on Π to points on Π' and then from points on Π' to points on Π''. It is easily shown (8.9 Ex. 25) that the transformations form a group under the operation of multiplication. The question is: for a projection of the group, what properties of figures remain invariant? The most striking is a property, the 'cross-ratio', of four points A, B, C and D on a line:

DEFINITION: *The* **cross-ratio** *of four collinear points*

$$(ABCD) = \frac{AB\,.\,CD}{AD\,.\,CB} = \frac{AB}{CB}\Big/\frac{AD}{CD}.$$

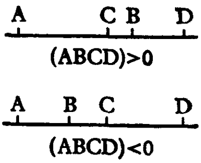


FIG. 8.7*b*

Here we can regard A and C as reference points in terms of which we can assess the position of two other points B and D. One ratio AB/CB fixes B in relation to A and C, where AB and CB are lengths with an appropriate sign attached. In Fig. 8.7*b*, positive distances are to the right, negative to the left. So AB and CB are both positive if B is to the right of A and $C\,(AB/CB > 0)$ but AB is positive and CB negative if B is between A and $C\,(AB/CB < 0)$. Similarly, the other ratio AD/CD fixes D in relation to A and C. The cross-ratio is obtained by dividing one ratio AB/CB by the other AD/CD; it can be positive or negative. Two cases where $(ABCD)>0$ and $(ABCD)<0$ respectively are illustrated; other cases can be drawn with the four points in various positions. It is easily checked that $(ABCD)>0$ means that both or neither of B and D lie between A and C, $(ABCD)<0$ that just one of B and D lies between A and C. The particular case $(ABCD)=-1$ has B and D dividing AC in the same ratio, one internally and the other externally (8.9 Ex. 28).

From a given centre Q, project collinear points $ABCD$ on Π into collinear points $A'B'C'D'$ on Π'. All eight points and Q lie on one plane, cutting Π in the line λ and Π' in the line λ'. Fig. 8.7c is drawn in this plane. The invariant property of projective geometry is:

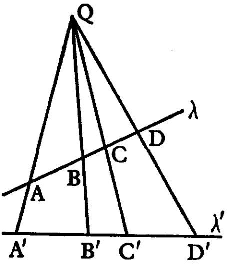


Fig. 8.7c

Theorem: *The cross-ratio of four points is invariant under projection:*

$$(ABCD)=(A'B'C'D').$$

To prove, use formulae for the area of a triangle (Appendix A.8):

$$\frac{1}{2}\,QA\,.\,QB\sin\angle AQB=\text{Area } QAB=\frac{h}{2}\,AB$$

where h is the length of the perpendicular from Q to the line λ. So:

$$AB=\frac{1}{h}\,QA\,.\,QB\sin\angle AQB \quad \text{and similarly for other lengths.}$$

$$\text{Hence: } (ABCD)=\frac{\frac{1}{h}\,QA\,.\,QB\sin\angle AQB\times\frac{1}{h}\,QC\,.\,QD\sin\angle CQD}{\frac{1}{h}\,QA\,.\,QD\sin\angle AQD\times\frac{1}{h}\,QC\,.\,QB\sin\angle CQB}$$

$$=\frac{\sin\angle AQB\,.\,\sin\angle CQD}{\sin\angle AQD\,.\,\sin\angle CQB}$$

$$\text{Similarly: } (A'B'C'D')=\frac{\sin\angle A'QB'\,.\,\sin\angle C'QD'}{\sin\angle A'QD'\,.\,\sin\angle C'QB'}$$

$$=\frac{\sin\angle AQB\,.\,\sin\angle CQD}{\sin\angle AQD\,.\,\sin\angle CQB}=(ABCD) \quad \text{Q.E.D.}$$

This is a very ıemarkable result. In a projection from Π to Π', the ratio AB/CB gets changed to $A'B'/C'B'$. At the same time the ratio AD/CD is changed to $A'D'/C'D'$. The invariance of the cross-ratio simply implies that, no matter what planes are used in the projection, the ratio AB/CB changes proportionately with the ratio AD/CD. For example, if one ratio is doubled, so is the other.

Projective geometry has to do with the art of perspective drawing. Moreover projective methods can be used to establish geometric properties (not depending on distances and angles), properties which are proved only with great difficulty by traditional methods. Consider, as a simple example, the 'complete quadrilateral' $ABCDEF$ shown in Fig. 8.7*d*. The diagonals AC, BD and EF meet in three points P, Q and R. The result is: P and Q divide the diagonal AC internally and externally in the same ratio, and similarly for the other two diagonals.

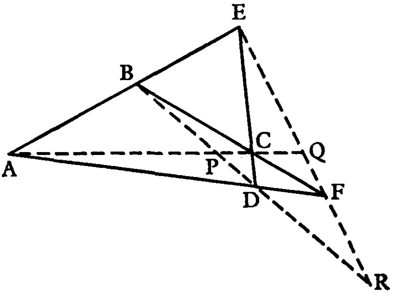


FIG. 8.7*d*

Proof: by projection from the line BD to the line EF, with centre at A, the cross-ratios $(BPDR)$ and $(EQFR)$ are equal. Similarly, by projection from BD to AC, with centre at E, the cross-ratios $(BPDR)$ and $(APCQ)$ are equal. Hence:

$$(APCQ)=(BPDR)=(EQFR)=\lambda \text{ (say).}$$

It remains to find λ. By projection from BD to AC, with centre at F, the cross-ratios $(BPDR)$ and $(CPAQ)$ are equal. Hence

$$(CPAQ)=\lambda=(APCQ).$$

So: $$\lambda^2=(APCQ)\times(CPAQ)=\frac{AP\,.\,CQ}{AQ\,.\,CP}\frac{CP\,.\,AQ}{CQ\,.\,AP}=1\,.$$

Hence $\lambda=\pm 1$ and, by the grouping of the points in $(APCQ)$, $\lambda=-1$. The cross-ratio $(APCQ)=-1$ on the diagonal AC. The case of cross-ratio equal to -1 has already been noted (8.9 Ex. 28 again); it implies that P and Q divides AC internally and externally in the same ratio. Similarly, $(BPDR)=-1$ and $(EQFR)=-1$ with the same result for the other two diagonals. Q.E.D.

8.8. Homogeneous co-ordinates. Consider the lack of symmetry between points and lines noted in 8.2. We say that any two points are joined by a unique line and we would like to say that any two lines meet in a unique point. We are held up by the difficulty about parallel lines. Again, consider the relation $ax+by+c=0$. If the ratios $a:b:c$ are given (which makes two parameters), the relation shows a variable point (x, y) on a fixed line. If the values of x and y are given (again two parameters), the relation shows a variable line (as a, b and c take different values) always passing through a fixed point (x, y). The relation has a dual interpretation, according as $a:b:c$ or x and y are given, i.e. it is a variable point on a fixed line or a variable line through a fixed point. The lack of symmetry arises since the co-ordinates of the fixed point are two *numbers* x and y whereas the 'co-ordinates' of the fixed line are two *ratios* of numbers $a:b:c$. Everything would be symmetric if the co-ordinates of a fixed point were (x, y, z), the ratios $x:y:z$ being given. The idea here is to represent a point on a plane by three co-ordinates rather than two, i.e. 'homogeneous co-ordinates' (x, y, z) considered as two ratios $x:y:z$. This is in the interests of symmetry, of conceptual neatness. Moreover, though it may seem confusing, it is not without practical use.

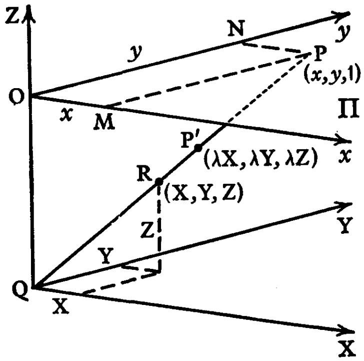


Fig. 8.8a

P is a point (x, y) referred to axes Oxy in a plane Π. In Fig. 8.8a, P is located by $OM=x$ and $ON=y$. Draw OZ at right angles to Π and take Q on it, unit distance below O. Draw QX and QY parallel to Ox and Oy respectively. Then, referred to axes $QXYZ$, any point R in three dimensions has co-ordinates (X, Y, Z). In particular, O is $(0, 0, 1)$ and P $(x, y, 1)$.

Consider any point P' on the line QR, with co-ordinates $(\lambda X, \lambda Y, \lambda Z)$ for various real values λ. As special cases, $\lambda=0$ gives Q, $\lambda=1$ gives R and $\lambda=1/Z$ gives P, provided that P is on QR: $x=X/Z$, $y=Y/Z$. Hence any point on QR can be assigned the *same* co-ordinates (X, Y, Z) as long as only the ratios $X:Y:Z$ are used, i.e. treating $(\lambda X, \lambda Y, \lambda Z)$ as (X, Y, Z). In particular, the point P

in Π has the co-ordinates (X, Y, Z). This means that its actual co-ordinates in three dimensions are $(X/Z, Y/Z, 1)$ referred to $QXYZ$, and that its actual co-ordinates in the two dimensions of Π are (x, y) referred to Oxy, where $x=X/Z$ and $y=Y/Z$. With this convention, any projection from Q (e.g. of P into P') leaves the *homogeneous co-ordinates* (X, Y, Z) of a point unaltered. The actual co-ordinates in three dimensions are changed, $(\lambda X, \lambda Y, \lambda Z)$ for various λ, but this matters only if we are concerned with distances and angles.

In a plane Oxy, a point is (x, y) and a line is $ax+by+c=0$. To convert to homogeneous co-ordinates, write $x=X/Z$, $y=Y/Z$, denote the point by (X, Y, Z) and the line by $aX+bY+cZ=0$. There is now complete symmetry: a point (X, Y, Z) given by ratios $X : Y : Z$ and a line (a, b, c) given by ratios $a : b : c$. Moreover, in this form, nothing is changed by projection from the point Q from one plane to another.

We are now in a position to handle parallel lines. Consider a plane Π' cutting $QXYZ$ in A', B' and C' respectively (Fig. 8.8.*b*). Project from Π' onto Π with centre Q so that P' goes into P, (X, Y, Z) being the homogeneous co-ordinates of each. Then C' goes into O, both with co-ordinates $(0, 0, Z)$. The line $C'A'$ goes into Ox, both represented by $Y=0$. The line $C'B'$ goes into Oy, or $X=0$. What of A', B' and the line $A'B'$? A' has co-ordinates $(X, 0, 0)$ and so has its image in Π. Since QA' is parallel to Π, there is no (finite) point of intersection. But we have a 'point' specified by co-ordinates $(X, 0, 0)$. By convention, label this point in Π as the 'point at infinity' on Ox, co-ordinates $(X, 0, 0)$. Similarly, B' goes into the 'point at infinity' on Oy, $(0, Y, 0)$. Finally, the line $A'B'$ is given by $Z=0$ and this goes into the 'line at infinity' on Π, equation $Z=0$.

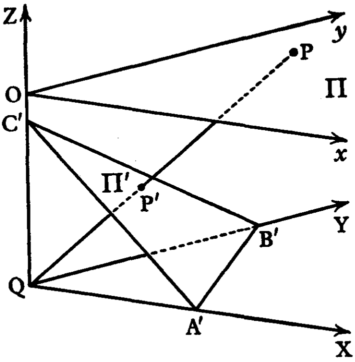


FIG. 8.8*b*

Homogeneous co-ordinates (X, Y, Z) for points in a plane Π therefore work as follows. If $Z\neq 0$, we have a (finite) point in the plane, with homogeneous co-ordinates (X, Y, Z) or $(X/Z, Y/Z, 1)$.

The actual co-ordinates referred to Oxy are $x = X/Z$, $y = Y/Z$. If $Z = 0$, the point $(X, Y, 0)$ is a point at infinity, e.g. $(X, 0, 0)$ at infinity on Ox and $(0, Y, 0)$ at infinity on Oy. In terms of lines, $aX + bY + cZ = 0$ is any line in the plane Π in homogeneous co-ordinates. As special cases, $X = 0$ is recognised as the line Ox and $Y = 0$ as the line Oy. Further, $Z = 0$ is now to be interpreted as the line at infinity. In short, to get a point at infinity, on the line at infinity, simply write $Z = 0$. The actual co-ordinates (X/Z and Y/Z) of a point referred to Oxy have no meaning for points at infinity; but the homogeneous co-ordinates can be written.

To pursue the symmetry between points and lines, consider the dual cases:

(i) Given two points $P_1(X_1, Y_1, Z_1)$ and $P_2(X_2, Y_2, Z_2)$, find the line joining them, i.e. find $a : b : c$ so that $aX + bY + cZ = 0$ goes through P_1 and P_2:

$$aX_1 + bY_1 + cZ_1 = 0 \quad \text{and} \quad aX_2 + bY_2 + cZ_2 = 0$$

giving:
$$\frac{a}{Y_1Z_2 - Y_2Z_1} = \frac{b}{Z_1X_2 - Z_2X_1} = \frac{c}{X_1Y_2 - X_2Y_1} \quad \text{.........(1)}$$

The denominators in (1) are all zero only if $X_1 : Y_1 : Z_1$ are the same ratios as $X_2 : Y_2 : Z_2$, i.e. only if P_1 and P_2 are the same point. If P_1 and P_2 are distinct, then there are *always* ratios $a : b : c$ for the line P_1P_2, given by (1). This is true of points at infinity, as well as finite points. Write $Z_1 = Z_2 = 0$ (P_1 and P_2 points at infinity) and (1) gives $a = b = 0$, i.e. the line P_1P_2 is $Z = 0$. The line joining any two points at infinity is the line at infinity.

(ii) Given two lines

$$\lambda_1(a_1X + b_1Y + c_1Z = 0) \quad \text{and} \quad \lambda_2(c_2X + b_2Y + c_2Z = 0),$$

find their point of intersection, i.e. find $X : Y : Z$ so that both equations hold:

$$a_1X + b_1Y + c_1Z = 0 \quad \text{and} \quad a_2X + B_2Y + c_2Z = 0$$

giving:
$$\frac{X}{b_1c_2 - b_2c_1} = \frac{Y}{c_1a_2 - c_2a_1} = \frac{Z}{a_1b_2 - a_2b_1} \quad \text{.......................(2)}$$

which is the dual of (1). Again, if the lines λ_1 and λ_2 are distinct ($a_1 : b_1 : c_1$ different from $a_2 : b_2 : c_2$), then there are *always* ratios $X : Y : Z$ for the point of intersection, given by (2). This is true of parallel lines as well as lines which are not parallel. Write $\frac{a_1}{b_1}=\frac{a_2}{b_2}$ (λ_1 and λ_2 parallel) and (2) gives $Z=0$, i.e. a point at infinity. The point of intersection of parallel lines is a point at infinity. The dual is complete; points at infinity are 'where parallel lines meet'.

Further tidying up is now possible. Two lines meet in one point since there is a single (unique) solution of two (distinct) linear equations. What of loci given by relations of second or higher degree? The number of points of intersection should clearly be dictated by the number of roots of the corresponding equations. Let us explore this question for the circle, one of the curves with a second-degree equation. Given two second-degree equations in x and y, the elimination of y generally produces an equation in x of the fourth degree with four roots, i.e. four real roots, *or* two real roots and a conjugate complex pair, *or* two conjugate complex pairs. Hence two circles should meet in four points: four real points, *or* two real points and two 'imaginary' points, *or* four 'imaginary' points. This doesn't seem to fit with facts, for no pair of circles appears to give four points of intersection; two points of intersection (real or 'imaginary') are all we can locate.* What of the missing pair? They can be located by use of homogeneous co-ordinates.

The equation of a circle with given centre and radius (8.6) is

$$(x-a)^2+(y-b)^2=r^2$$

or

$$(X-aZ)^2+(Y-bZ)^2=r^2Z^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots(3)$$

in homogeneous co-ordinates. The line at infinity $Z=0$ intersects the circle in two points $(X, Y, 0)$ where $X : Y$ is given by (3) on putting $Z=0$:

$$X^2+Y^2=0 \quad \text{or} \quad Y=\pm iX.$$

Since only the ratio matters, write $X=1$, $Y=\pm i$. Hence the circle (3) cuts the line at infinity in the two points $(1, i, 0)$ and $(1, -i, 0)$. Any line cuts a circle in two points, real or 'imaginary' (8.9 Ex. 32). We have found the two points where the circle cuts the line at infinity; they are 'imaginary' as well as on the line at infinity. What is so remarkable, however, is that the points $(1, \pm i, 0)$ do not depend on a, b and r, the parameters in (3) which locate the centre and radius.

* Compare the case of two ellipses, also with second-degree equations, where *four* points of intersection (real or imaginary) can be located. See 8.9 Ex. 34.

All circles go through the *same* pair of points (1, $\pm i$, 0), called the *circular points at infinity*. The answer to our question is now clear (and further illustrated in 8.9 Ex. 33). Two circles do intersect in four points, i.e. the two (imaginary) circular points at infinity, together with two other points, real or imaginary.

8.9. Exercises

1. *Rigid motions.* Show that a horizontal translation of a figure in a plane (as in Fig. 6.3*c*) can be achieved by two reflections, by turning the figure over a vertical line and over again.

2. Show that the result of Ex. 1 holds for a rotation about a fixed point O, by turning the figure over and over again about lines through O. Generalise for combinations of translations and rotations and hence for all rigid motions. Distinguish between the result of an even number and an odd number of reflections.

3. *Gradients.* In a triangle OAB, take OA horizontal and OB rising (e.g. a road on a hill). Interpret the statement that 'the gradient of OB is 1 in 12' and indicate that there may be some uncertainty. Suggest a strict definition of *gradient* as the *slope* of a line (8.6).

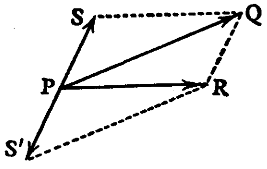


FIG. 8.9*a*

4. Demonstrate the validity of the equation of PR to $PQ - PS = PQ + (-PS)$ in Fig. 8.9*a* by interpreting $(-PS)$ as PS'.

5. *Vectors as an additive group.* From the properties (ii) of 8.3, check that the set of all vectors is a commutative group under addition, with identity the zero vector.

6. In a vector space V (over F), show that $0u = 0$ for all vectors u. Proceed: write $u + 0u = 1u + 0u = (1+0)u$ by rules S2 and S4. Use the facts that 0 in F is such that $1 + 0 = 1$ and 0 in V is such that $u + 0 = u$, and the cancellation rule for the additive group of vectors.

7. By a method similar to Ex. 6, show that $a0 = 0$ for all scalars a, where 0 is the zero vector. Compare with $a0 = 0$ for a field (0 the zero scalar) as in 6.9 Ex. 20

8. The additive group of the vector space V has both u and its negative $-u$. Proceed: $u + (-u) = 0 = 0u = \{1 + (-1)\}u$, and reduce by use of S2 and S4. Hence establish that $(-1)u = -u$.

9. *Euclidean space: sum as resultant of vectors.* The sum of $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ is defined as $(x_1 + x_2, y_1 + y_2)$. Let P complete the parallelogram as in Fig. 8.9*b*. Show $M_2M = x_1$, $NP = y_1$ and hence $OM = x_1 + x_2$, $MP = y_1 + y_2$. Deduce that P is the sum of P_1 and P_2, OP being the resultant of OP_1 and OP_2. Extend to Euclidean space of 3 dimensions.

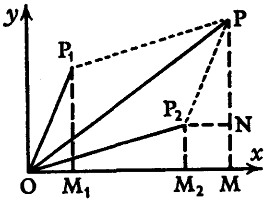


FIG. 8.9*b*

10. From the definition of the sum of (x_1, y_1) and (x_2, y_2) as (x_1+x_2, y_1+y_2), show that the negative of (x, y) in the additive group of vectors is $(-x, -y)$ and that the difference: (x_2, y_2) less (x_1, y_1) is (x_2-x_1, y_2-y_1). Show the point P' (x_2-x_1, y_2-y_1) graphically relative to P_1 (x_1, y_1) and P_2 (x_2, y_2) and that OP_2 is the resultant of OP_1 and OP'.

11. Develop the geometric concept of a vector PQ as the *relative* location of two points P and Q (with no fixed position) by taking an origin O in the plane and showing that the *vector* PQ is equivalent to OR, i.e. to the *point* R once O is fixed (Fig. 8.9c). Further, if axes Oxy are selected to give co-ordinates $P(x_1, y_1)$ and $Q(x_2, y_2)$, then the vector PQ is equivalent to the point R (x_2-x_1, y_2-y_1). Hence show there is no inconsistency in taking an algebraic vector (x, y) either as a geometric point (x, y) or as a geometric vector PQ, where x and y are the *differences* of the co-ordinates of P and Q.

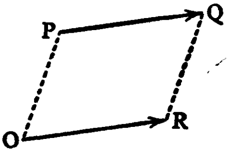


FIG. 8.9c

12. Follow up Ex. 11 by showing that the *distance* between two vectors (e.g. PQ of Fig. 8.9c) is the same concept as the *length* of one vector (e.g. OR of Fig. 8.9c).

13. *Lines defined by scalar products.* $P(x_1, y_1)$ and $Q(x_2, y_2)$ are two fixed points referred to axes Oxy. If $S(x, y)$ divides PQ in the ratio $\lambda : 1-\lambda$ where $0 \leqslant \lambda \leqslant 1$, show that $x=(1-\lambda)x_1+\lambda x_2$ and $y=(1-\lambda)y_1+\lambda y_2$. Check for $\lambda=0, \frac{1}{2}, 1$. Use R of Ex. 11 and take T on OR in ratio $\lambda : 1-\lambda$. Show that T has co-ordinates $\lambda(x_2-x_1)$ and $\lambda(y_2-y_1)$, i.e. T is the scalar product of R for variable scalars λ. Hence show that the line PQ can be defined by varying S corresponding to varying T, i.e. by varying scalar products. Deduce its equation:

$$\frac{x-x_1}{x_2-x_1}=\frac{y-y_1}{y_2-y_1},$$

with common value λ. See (3) of 8.6. Is the restriction $0 \leqslant \lambda \leqslant 1$ necessary?

14. *Mapping the vectors* $(-x, y)$. Show that the mapping of the vectors $(-x, y)$ for x and y positive, is a reflection in Oy of the mapping of the vectors (x, y) (Fig. 8.9d). Establish that the mapping is completed once the point A' for the vector $(-1, 0)$ is inserted and that the vector OA' is the negative of OA, i.e. the product by the scalar -1.

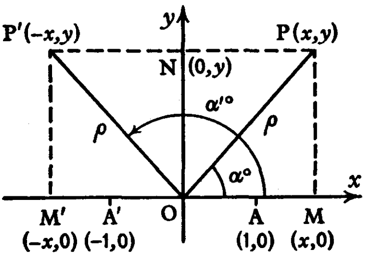


FIG. 8.9d

15. In Fig. 8.9d, take $\alpha'=180°-\alpha$. Extend the relations

$$x=\rho\cos\alpha \quad \text{and} \quad y=\rho\sin\alpha$$

for $P(x, y)$, where $x>0$ and $y>0$, so that they apply to $P'(-x, y)$. Show that $\cos\alpha'$ must be written as OM'/OP' and $\sin\alpha'$ as $M'P'/OP'$ and that this is equivalent to defining $\cos(180°-\alpha)=-\cos\alpha$ and $\sin(180°-\alpha)=\sin\alpha$ (Appendix A.7).

16. Complete the extension of Ex. 14 and 15 to accommodate vectors (x, y) for all real x and y and trigonometric ratios of angles α where $0° \leqslant \alpha \leqslant 360°$.

17. A line has slope m and intercept c on Oy. From (3) of 8.6, noting that the line goes through $(0, c)$, show that it has equation $y = mx + c$. Conversely, show that the locus $y = mx + c$ is a line of slope m and intercept c on Oy.

18. Show that, if (x_1, y_1) and (x_2, y_2) are points on a line parallel to Ox, then $y_1 = y_2$. Deduce that the line has equation $by + c = 0$, or $y =$ constant. Compare with the line $x =$ constant parallel to Oy. Check from formula (3) of 8.6.

19. The line parallel to Ox has slope zero; what can be said of the slope of the line parallel to Oy?

20. Show that $x + y = 1$ is the line through $(1, 0)$ and $(0, 1)$. Generalise and establish that $bx + ay = ab$ is the line with intercepts a and b on the axes.

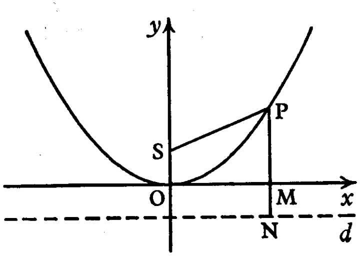


FIG. 8.8*e*

* 21. *The parabola.* S is a given point, distant 2α from a given line d. A point P moves so that the distance SP is always equal to the perpendicular distance PN from d. The locus P is described as a *parabola* with *focus* S. Fix axes so that S is $(0, \alpha)$ and d is $y = -\alpha$ (Fig. 8.9*e*) and show that the equation of the locus is $y = \frac{1}{4\alpha} x^2$. (Write expressions for SP and PN in terms of $OM = x$, $MP = y$, and equate.)

* 22. Conversely, show that $y = ax^2$ $(a > 0)$ is a parabola with focus at $S\left(0, \frac{1}{4a}\right)$. Show also that $y = ax^2$ $(a < 0)$ is a parabola, the reflection of the locus of Fig. 8.9*e* in Ox. This latter parabola can represent the path of a ball thrown into the air (neglecting air resistance).

23. *The group of rigid motions.* A transformation sends $P(x, y)$ into $P'(x', y')$ by reflection in Ox; show that $x' = x$ and $y' = -y$ and that the distance between two points is unchanged by the transformation. Use Ex. 2 to deduce that the same is true of all transformations of the group of rigid motions.

24. Use (2) of 8.4 to show that the angle between two lines (vectors) is invariant under a transformation of the group of rigid motions.

25. *The group of projections.* A figure is projected (from a fixed point Q) from a plane Π to a plane Π'. Check that the set of all such projections, for various Π and Π', forms a group, the identity being the projection from Π to Π and the inverse of the projection from Π to Π' being that from Π' to Π.

26. Points A, B, C and D are on a line λ at distances a, b, c and d respectively from a fixed point on λ. Show that the cross-ratio $(ABCD) = \frac{(b-a)(d-c)}{(d-a)(b-c)}$.

27. Keep A fixed and permute B, C and D in $(ABCD)$. If one of the six cross-ratios, $(ABCD)$, $(ABDC)$, ..., has value μ show that the others are $1/\mu$,

$1-\mu, \dfrac{1}{1-\mu}, 1-\dfrac{1}{\mu}$ and $1\Big/1-\dfrac{1}{\mu}$. Check by writing $(ABCD)=\dfrac{b(d-c)}{d(b-c)}$ (Ex. 26, case $a=0$) and by permuting b, c and d.

28. *Harmonic division.* Show that $(ABCD)=-1$ is equivalent to

$$AB/BC=AD/CD,$$

i.e. B and D divide the segment AC internally and externally in the same ratio.

* 29. *Pencils of lines.* A set of lines $\{a, b, c, \ldots\}$ passing through a fixed point P form a *pencil* (Fig. 8.9*f*). If (ab) denotes the angle between lines a and b, define the *cross-ratio* of a pencil of four lines as $\dfrac{\sin(ab)\sin(cd)}{\sin(ad)\sin(cb)}$ and show that it equals $(ABCD)$ where A, B, C and D are the intercepts of the pencil and a line λ not through P. Deduce that this cross-ratio is also invariant under projections. Comment on the duality displayed between points and lines.

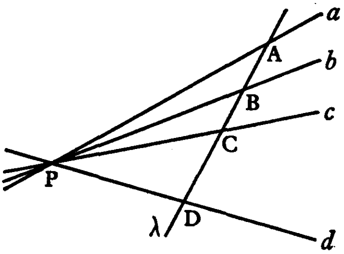


FIG. 8.8*f*

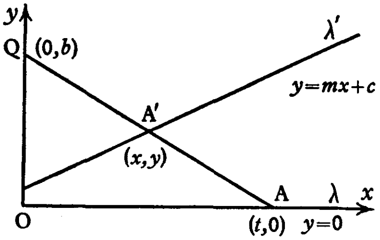


FIG. 8.8*g*

* 30. In projecting (from a fixed point Q) from line λ to line λ', choose axes as shown in Fig. 8.9*g*. The point A' (x, y) on λ' corresponds to a point A $(t, 0)$ on λ. By solving the equations for the line QA and that for λ' $(y=mx+c)$, show that $x=(b-c)t/(b+mt)$. Deduce that the transformation of such a projection can always be written algebraically as $x=\alpha t/(\beta+\gamma t)$ for parameters α, β and γ.

31. In homogeneous co-ordinates, show that the line joining O $(0, 0, 1)$ to a fixed point $P(X_1, Y_1, Z_1)$ is $aX+bY=0$ where $a/b=-Y_1/X_1$. Similarly show that the line joining $(1, 0, 0)$ and (X_1, Y_1, Z_1) is $bY+cZ=0$ where $b/c=-Z_1/Y_1$. Interpret as a line parallel to Ox.

32. A circle and a line are given; write their equations: $x^2+y^2=r^2$ and $ax+by+c=0$ (the origin being at the centre of the circle). Show that there are two points of intersection. For what values of $a:b:c$ are they imaginary?

33. Illustrate that two circles cut in only two points (real, coincident or imaginary) by finding the intersections of $x^2+y^2=1$ with each of

$$(x-1)^2+y^2=1, \quad (x-2)^2+y^2=1 \quad \text{and} \quad (x-3)^2+y^2=1.$$

Illustrate graphically. Put into homogeneous co-ordinates and find in each case the circular points at infinity $(1, \pm i, 0)$ as intersections.

*34. *The ellipse.* An ellipse with fixed axes (taken as Ox and Oy) is the locus with equation $x^2/a^2+y^2/b^2=1$ for parameters a and b. Solve $3x^2+y^2=1$ and $x^2+3y^2=1$ and illustrate that two such ellipses can intersect in four real points, here $(\pm\frac{1}{2}, \pm\frac{1}{2})$. Show that nothing is added by putting the equations into homogeneous co-ordinates.

*35. *Conic sections.* A circle C is given in the plane Π. Project from a point Q outside Π onto another plane Π'. Show that C is sent into a circle, ellipse, parabola or hyperbola according to the position of Π'. These are 'conic sections', i.e. plane sections of the cone with vertex Q and base C.

*36. *Affine geometry.* As a special case of projective geometry, project from a plane Π to a plane Π' by parallel lines. If Π and Π' are not parallel planes, show that a circle in Π is sent into an ellipse in Π'. This is 'affine geometry' and the corresponding transformations are those of the 'affine group' (7.9, Ex. 24).

CHAPTER 9

LIMITS AND CONTINUITY

9.1. Functions of a real variable. Before embarking on an exploration of new territory, that of mathematical analysis or the calculus, we can profitably pause to take stock of our equipment. Analysis can be regarded as a very specialised and highly elaborate extension of algebra, developed from two particular concepts: the real number system and the functional relation. These are somewhat incidental in algebra but they form the essential foundation on which the powerful techniques of the calculus rest.

In algebra we are concerned with sets of elements which may or may not be numbers. Even when we deal with numbers, we are often quite happy to stick to rational numbers, making up an ordered field with the long list of properties of 2.2. It is true that purely algebraic considerations require an extension of the number system to real and complex numbers; without them we have no zeros of polynomials as simple as x^2-2 or x^2+2. But the exploitation of real numbers is the job we undertake in analysis rather than in algebra. The property to exploit is that real numbers form a *complete* ordered field (2.4 and 6.7):

> If a set of real numbers has a lower bound, then it has a GLB; if a set has an upper bound, then it has a LUB.

This is the feature which distinguishes the real numbers from the integers and rationals. A set of numbers x such that $x^2<2$ (or equally $x^2\leqslant 2$)* is bounded in both directions; it has no LUB or GLB if x is confined to the rationals whereas it has LUB $=\sqrt{2}$ and GLB $=-\sqrt{2}$ if x is a real number.

In algebra, also, we are interested in relations between two sets

* There is no difference between the two sets if x is rational. The difference when x is real is that the set such that $x^2<2$ does not include its LUB $\sqrt{2}$ whereas the other set does (and similarly for the GLB $=-\sqrt{2}$).

and, rather incidentally, with the particular kind of relation called a function (7.3). In analysis, however, the particular case is the one pursued. Analysis deals with sets of real numbers and with functional relations between them; it is the study of real-valued functions of a real variable. It is important to be quite explicit on this limitation. A *relation* R is any subset of the Cartesian product $X \,.\, Y$ of two sets X and Y of any elements, i.e. a set of ordered pairs

$$\{(x, y) \mid x \in X, y \in Y, yRx\},$$

where yRx is a statement linking x and y. The relation is a *function* when the statement is such that, if $x \in X$ has a corresponding $y \in Y$, then y is unique. Let X be limited to the *domain* of the function, i.e. the set of first elements x of the pairs (x, y) which correspond to some y; let Y be limited to the *range* of the function, i.e. the set of second elements y of the pairs (x, y) which correspond to some x. Then a function is a many-one mapping f of the set X onto the set Y. The rule which specifies which y corresponds to a given x is a statement yfx, usually written $y=f(x)$. There is, as yet, no limitation on the kind of elements comprised in the sets X and Y. Now suppose that X and Y are sets of real numbers, where $x \in X$ and $y \in Y$ are called *variables*, as opposed to single real numbers which are *constants*. To distinguish the variables of a function, x in the domain X of the function is called the *independent variable* and y in the range Y of the function is the *dependent variable*. We have then a *real-valued function of a real variable*, a many-one mapping of one set X of real numbers (the domain of the independent variable x) onto another set Y of real numbers (the range of the dependent variable y).

There are always two things to specify for a function of a real variable:

(i) the sets of real numbers involved in the many-one mapping $X \rightarrow Y$, i.e. the domain X and the range Y of the function, and

(ii) the rule $y=f(x)$ of the mapping, i.e. how to get from a given x of the domain to the unique image y in the range.

Both are apparent in the full notation for a function as the set $\{(x, y) \mid x \in X, y \in Y, y=f(x)\}$. Both are stressed in the shorter notation 'f: $X \rightarrow Y$', as sometimes used for a function having the rule f in the mapping of X onto Y.

In practice, we adopt a looser terminology. We have in mind the rule $y=f(x)$ of a function and (perhaps to a less extent) the domain X on which it is defined. The range Y is to be found by taking all possible x's in the domain X and seeing what y's correspond. Hence we speak of 'the function f whose values are $y=f(x)$ defined on the domain X'. Quite usually, we make no explicit reference to the domain X, speaking simply of 'the function $y=f(x)$'. This is indeed an elliptic expression since we are transferring the term 'function' from the mapping of one set onto another (which is what it is) to the particular rule $y=f(x)$ which specifies how y is obtained from x in the mapping. This is all very well, and a great saving of time, provided we remember what is implied, provided we leave no doubt about the domain of the function and about the consequent range of its values.*

In practice, also, the domain of a function is usually described loosely. Again there is no difficulty, provided we remember that the domain is a set of real numbers. So, if a function has domain $\{x \mid x \text{ a real number}\}$, we may say that the function is defined 'for all x'. Or, if the domain is the set of real numbers between 0 and 1 inclusive, i.e. the set $\{x \mid x \text{ a real number}, 0 \leqslant x \leqslant 1\}$, we may speak of the function as defined 'on $0 \leqslant x \leqslant 1$'.

For example, consider the linear function $y=2x-1$. Here we may mean that the set X of all real numbers is mapped onto the set Y of all real numbers by the linear rule $y=2x-1$, and it may be safe enough to leave the domain X understood. Or, we may have some other domain X in mind, e.g. the set $X=\{x \mid x \text{ a real number}, 0 \leqslant x \leqslant 1\}$, in which case we may amplify: the linear function $y=2x-1$ defined on $0 \leqslant x \leqslant 1$. Again, write the function $y=\sqrt{x}$. Here we must be more careful. The rule $y=\sqrt{x}$, of the mapping $X \rightarrow Y$ which is the function, is to be interpreted: given a real number x, then y is the value obtained as the positive square root of x. Hence, x must be zero or positive and y also zero or positive. The domain X of the function can *not* be the set of all real numbers. It *can* be the set of all non-negative real numbers or any subset such as all real $x>1$. We

* Note the definition given by Dirichlet (1805–59): $f(x)$ is a real function of a real variable if, to every real number x, there corresponds a real number $f(x)$. Here the second '$f(x)$' is the rule for getting from x in X to $y=f(x)$ in Y. Something needs to be specified about what set X of real numbers is the domain and hence about what set Y of real numbers is the range. The first '$f(x)$' is then to be interpreted as $f\colon X \rightarrow Y$.

must say what we have in mind, e.g. the function $y=\surd x$ defined on $x\geqslant 0$ (or on $x>1$, or whatever the domain may be).

In this chapter, the main development is in terms of a general function, and of its general properties, without specifying any particular rule. We need a general and flexible notation for the rule of the function, one which is capable of distinguishing many different functions. Ringing the changes on small and capital letters, and on the English and Greek alphabets, we can write:

$$f(x),\ g(x),\ \dots\ F(x),\ G(x),\ \dots\ \phi(x),\ \psi(x),\ \dots .$$

A notation such as $y=f(x)$ applies to the rule of the function; it needs to be completed by specification of the domain X. A different function arises if the rule is changed or if the domain is varied (or both). So $y=f(x)$ defined on X is a different function from $y=g(x)$ defined on X; different letters f and g indicate that the rule is changed. Equally, $y=f(x)$ defined on X_1 is a different function from $y=f(x)$ defined on X_2 since they have different domains X_1 and X_2; the same letter f is used when the rule for getting from x to y is not varied. For example, $y=x^2$ defined for all x, $y=x^2$ defined for $x>0$, $y=\surd x$ defined for $x>0$ are three different functions. The rule is the same for the first two so that $y=f(x)$ can be used for both, with $f(x)=x^2$ here. The rule is different for the last two so that, if $y=f(x)$ is used for one, then a different letter (say g) is needed in writing $y=g(x)$ for the other.

On the other hand, a change in the labels of the variables is purely formal, having no effect either on the function or (in particular) on the rule of the function. So $y=f(x)$, $x=f(y)$, $z=f(u)$, ... all represent the same functional rule and, if the domains (X, Y, U, ...) are the same sets of real numbers, they are all the same function. In the first we write x for the independent and y for the dependent variable, and we simply change these labels for the others. For example, $y=x^2$, $x=y^2$, $z=u^2$, ... defined on the domain (X, Y, U, ...) of all real numbers are all the same function.

9.2. Algebraic and other functions. In using particular functions for illustration, we draw upon rules obtained by algebraic operations. The following are some examples:

Rule $y=f(x)$	*Domain* X	*Rule* $y=f(x)$	*Domain* X
(i) $y=1+x+x^2$	all x	(ii) $y=\frac{1+x^2}{1-x^2}$	all $x (x \neq \pm 1)$
(iii) $y=\frac{x+\sqrt{x+1}}{\sqrt{x-1}}$	$x>1$	(iv) $y=\frac{1}{\sqrt{1-x^2}}$	$-1<x<1$

The importance of the specification of the domain is here further illustrated. In each case, the domain X shown is the widest set possible for y to be defined at all. Any subset of X can serve equally well as a domain. The range Y of each function follows from a consideration (sometimes quite involved) of the values y can take; as shown below, Y is all real $y \geqslant \frac{3}{4}$ for (i).

These are all cases of *algebraic functions*, i.e. the rules involve polynomials or root extraction (surds) and their ratios. They are called 'algebraic' since the rules are derived from x and given constants solely by repeated use of the operations of elementary algebra, i.e. the four rational operations $(+, -, \times, \div)$ together with the extraction of nth roots (n a positive integer).* The variety of elementary functions is extended very considerably later on (Chapter 12). Note, here, that a function of a real variable can be defined perfectly well by a rule which is not algebraic. Each of the following:

(v) $y=$ least integer not less than x

(vi) $y=1$ when x integral and $y=-1$ when x non-integral

(vii) $y=1$ when x rational and $y=-1$ when x irrational

satisfies all the conditions for a function of a real variable x, being a many-one mapping of the set X of all real numbers onto a set Y of real numbers. Y is the set of integers in the first case and the set $\{-1, 1\}$ in the others.

The geometric representation of functions, the basis of co-ordinate geometry (8.6), depends on the one-one correspondence between the ordered set X of real numbers and the set of points making up a directed line in space. This is an isomorphism preserving order; the complete ordering of the real numbers is matched by a complete ordering of points on a directed line. One representation (as in 7.3)

* Root extraction: $b=\sqrt[n]{a}$ implies $a=b^n$. It is the inverse of the process of multiplying a number by itself.

of a function $f: X \rightarrow Y$ is a mapping from points on one line to points on another line. More usually, the mapping of the function $y=f(x)$ proceeds by associating each pair (x, y) of the function with a point $P(x, y)$ referred to axes Oxy in a plane (as in 8.6). The result is that a function corresponds to a locus cut by lines parallel to Oy in no more than one point. An actual plotting on paper gives a graph of the function or locus. A graph of the function and locus (i) ($y=1+x+x^2$) is shown in Fig. 9.2, from which it is clear that the range of y is the set of all real $y \geqslant \frac{3}{4}$. In this case, the locus is recognisable as a 'curve' (a parabola). In other cases this is not so. An example is the graph of the function (vi). Fig. 9.2 shows that the locus here is the set of points on the line $y=-1$, except that a (countably infinite) number of points is missing, being replaced by points on the line $y=1$.

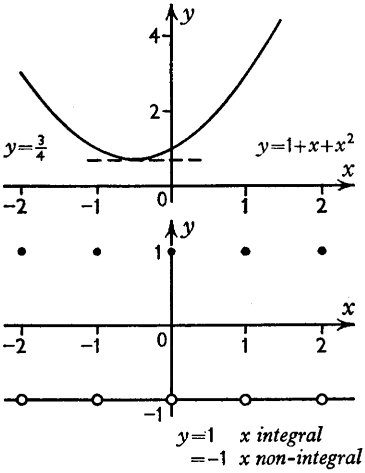


Fig. 9.2

9.3. The algebra of functions. A number of preliminary definitions and distinctions need to be made. The function $y=f(x)$ is defined on the domain X. Often X is the set of all real numbers $\{x \mid x \text{ a real number}\}$ or some 'half' set such as that of all positive real numbers $\{x \mid x \text{ a real number}, x>0\}$. Such domains correspond to the whole directed line Ox or to some 'half-line' of Ox. On other occasions it is either necessary, or at least useful, to limit X to an 'interval' of real numbers between specified values a and b, corresponding to a segment of the directed line Ox. In addition, however we define X, we often need to consider a 'neighbourhood' of real numbers around a particular value $x=\alpha$ within the domain X. Formally:

Definition: *An* **interval** $[a, b]$ *is the set* $\{x \mid x$ *a real number,* $a \leqslant x \leqslant b\}$ *for given real numbers a and b* ($a \leqslant b$). *A* **neighbourhood** N *of the real number* α *is an interval* $[a, b]$ *containing* α *within it* ($a<\alpha<b$).

An interval, often written shortly $a \leqslant x \leqslant b$, is a *closed* set of numbers containing both end values a and b.* A neighbourhood is defined broadly, as *any* interval containing α within it. Later, the idea is to concentrate on 'small' neighbourhoods or on neighbourhoods which get 'smaller'; this is something which needs to be developed carefully and in connection with the concept of a limit. Further, an interval or neighbourhood of values of x is specified so that corresponding values of $y=f(x)$ can be examined. Various questions arise. If $y=f(x)$ is defined on $[a, b]$, are the values of y bounded or not? If bounded, do the values of y themselves constitute an interval? Does a neighbourhood N of $x=\alpha$ correspond to a neighbourhood of values of y around $f(\alpha)$? These are by no means trivial questions; they need to be examined very closely.

Given various functions $f(x)$, $g(x)$, ..., we can write other functions by combining them by the operations of elementary algebra. Simple *combinations* are:

$$f(x)+g(x);\ f(x)-g(x);\ f(x)\times g(x);\ f(x)/g(x) \quad \text{for } g(x)\neq 0.$$

In such ways, complicated functions can be split into simpler ones, or given functions can be used to define new ones. It seems obvious enough but we must say exactly what we mean about the domains on which the combination functions are defined. Consider the sum function and suppose that $y_1=f(x)$ is defined on X_1, $y_2=g(x)$ on X_2. Both functions are defined on the set of x's common to X_1 and X_2, i.e. on $X=X_1 \cap X_2$. If $x \in X$, both y_1 and y_2 are uniquely defined and so is $y=y_1+y_2$. Hence we have the function $f(x)+g(x)$ defined on X, the intersection of the domains of the separate functions. The domain is the same for the other combinations, except that the ratio $f(x)/g(x)$ needs a qualification: it is defined on the domain $X=X_1 \cap X_1$ *except* that any α for which $g(\alpha)=0$ is excluded.

A more sophisticated, and very useful, combination of two given functions is the *composite function* or *function of a function*:

$$\text{Given } y=F(u) \text{ and } u=f(x), \text{ then } y=F\{f(x)\}.$$

* Variants can also be defined, open at one end or the other. So $a \leqslant x < b$ can be denoted as $[a, b[$, $a < x \leqslant b$ as $]a, b]$, $a < x < b$ as $]a, b[$. The conventional symbols '∞' and '$-\infty$' (read plus or minus 'infinity') appear in connection with limits below. By using them, it is possible to describe complete or 'half' sets of real numbers as intervals, e.g. $[a, \infty]$ for $x \geqslant a$, $[-\infty, b]$ for $x \leqslant b$, and $[-\infty, \infty]$ for all x. The minor convenience of such conventional notations is probably out-weighed by the dangers of using them; there is the temptation to write the interval $[a, b]$ and then to put $b=\infty$ (which is without meaning).

Here it is a matter of matching the *range* of $u=f(x)$ with the *domain* of $y=F(u)$ before we can specify a domain for $y=F\{f(x)\}$ as a function of x. To illustrate, consider the following cases where the domain X_1 of $u=f(x)$ is 'all x':

	$u=f(x)$	*Range* U_1	$y=F(u)$	*Domain* U_2	$y=F\{f(x)\}$	*Domain* X
(i)	$u=1+x$	all u	$y=u^2$	all u	$y=(1+x)^2$	all x
(ii)	$u=1+x+x^2$	$u\geqslant\frac{3}{4}$	$y=1/u$	all $u(u\neq 0)$	$y=1/1+x+x^2$	all x
(iii)	$u=1-x^2$	$u\leqslant 1$	$y=1/\sqrt{u}$	$u>0$	$y=1/\sqrt{1-x^2}$	$-1<x<1$
(iv)	$u=1-\sqrt{1+x^2}$	$u\leqslant 0$	$y=1/\sqrt{u}$	$u>0$	$y=1/\sqrt{1-\sqrt{1+x^2}}$	no x

Let $u=f(x)$ be defined on domain X_1 and have range U_1; let $y=F(u)$ be defined on domain U_2. In many cases, the range U_1 is either the same or included within the domain U_2, as illustrated by (i) and (ii). The domain X_1 of $f(x)$ then goes through to serve as the domain X of the composite function, i.e. $y=F\{f(x)\}$ is defined for the same values of x as $f(x)$. In other cases, the range U_1 and the domain U_2 are overlapping sets, as illustrated by (iii). The composite function is then defined only for those values of u which belong to both U_1 and U_2 (for $u\in U_1\cap U_2$). The domain X_1 (i.e. x giving $u\in U_1$) must be restricted to X (i.e. x giving $u\in U_1\cap U_2$) before it can serve as the domain of the composite function. So $y=F\{f(x)\}$ is defined for a smaller set of values of x than $f(x)$. In (iii), we start with the set of all x (giving $u\leqslant 1$) but we must restrict to $-1<x<1$ to ensure that $u>0$ (as well as $u\leqslant 1$) and that $1/\sqrt{u}$ is defined. Finally, as (iv) illustrates, it is quite possible that the range U_1 and domain U_2 do not overlap at all, so that the composite function is defined nowhere.

The function $y=f(x)$ is an 'increasing' one if, whenever we increase x from a to b, we also increase $f(x)$ from $f(a)$ to $f(b)$. A similar property is required of a 'decreasing' function. Both are very special cases of functions.*

DEFINITION: *$y=f(x)$ is an* **increasing function** *on the domain X if $a<b$ implies $f(a)<f(b)$ for all a and b in X; it is a* **decreasing function** *if $a<b$ implies $f(a)>f(b)$ for all a and b in X.*

In Fig. 9.3, cases (i) and (ii) illustrate increasing functions; cases (iii) and (iv) are functions neither increasing nor decreasing. The functions are here defined on the interval $[\alpha, \beta]$, or on some subset of the interval.

* A function which is *either* an increasing *or* a decreasing function on X is often described as a *monotonic* function, i.e. monotone increasing or monotone decreasing.

As a final distinction, we can enquire when a given function $y=f(x)$ also provides an *inverse function* $x=g(y)$. The given function is a many-one mapping $f\colon X\rightarrow Y$. Looking at the mapping the other way round, we expect in general that, to a specified y in Y, there correspond many x's in X. Hence, in general, x is *not* a function of y at all.

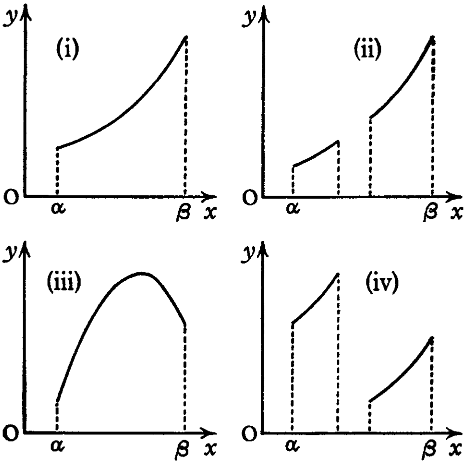


FIG. 9.3

The particular case where x *is* a function of y arises when the mapping $f\colon X\rightarrow Y$ is one-one. The inverse function $x=g(y)$ then exists and it can be denoted $x=f^{-1}(y)$, i.e. $f\colon X\leftrightarrow Y$ and $f^{-1}\colon Y\leftrightarrow X$ are the same mapping.

THEOREM: *A relation between sets X and Y gives both a function $y=f(x)$ and an inverse function $x=f^{-1}(y)$ if and only if the mapping $X\leftrightarrow Y$ is one-one.*

The notation, as so often in writing functions, tends to be somewhat confusing here. The function f and its inverse f^{-1} must not be taken as reciprocals: f^{-1} is not $\frac{1}{f}$; this is clear when we try to write the inverse of $y=f(x)$ as $x=\frac{1}{f}(y)$, a meaningless notation. Another confusion may arise from the practice of switching the labels of variables. Since $x=f(y)$ is the same function as $y=f(x)$, with variables interchanged, the same applies to $x=f^{-1}(y)$ and $y=f^{-1}(x)$. We can write $y=f(x)$ and $y=f^{-1}(x)$ as two different functions, one of which happens to be the inverse of the other. Then $y=f(x)$ implies $x=f^{-1}(y)$, sticking to the same labels for the variables. Equally, $y=f^{-1}(x)$ implies $x=f(y)$, again with the variables having unchanged labels. The labels are switched in passing from one statement to the other.

For example, take X and Y both as the set of non-negative real numbers. Then $y=x^2$ and $x=\surd y$ are alternative ways of writing the same one-one mapping between X and Y. Hence, $y=x^2$ and $y=\surd x$ are different functions, but one is the inverse of the other. Keeping to the same labels for the variables, we say that $y=x^2$ implies $x=\surd y$. Equally, we say that $y=\surd x$ implies $x=y^2$.

As another case (9.9 Ex. 10), consider the relation $x^2+y^2=1$ shown by a circle of unit radius, centred at O. Hence $y=\pm\surd(1-x^2)$ for values $-1\leqslant x\leqslant 1$; this is two-valued and not a function. If we take only the positive root, then $y=\surd(1-x^2)$ is a function defined on $-1\leqslant x\leqslant 1$, i.e. the semi-circle above Ox. On inversion, $x=\pm\surd(1-y^2)$ which is two-valued and not a function. Hence $y=\surd(1-x^2)$ as a function defined on $-1\leqslant x\leqslant 1$ does not have an inverse. Now take $y=\surd(1-x^2)$ defined on $0\leqslant x\leqslant 1$, i.e. a function represented by the quarter-circle in the positive quadrant. Then $x=\surd(1-y^2)$ defined on $0\leqslant y\leqslant 1$. This is also a function and it is the inverse of $y=\surd(1-x^2)$ defined on $0\leqslant x\leqslant 1$. It happens that $y=\surd(1-x^2)$ is its own inverse when both variables are restricted to the interval $[0, 1]$.

There is a connection between functions which are increasing (decreasing) and functions which possess inverses. Fig. 9.3 illustrates. An increasing function implies a one-one mapping; it has an inverse, which is also an increasing function. A similar result holds for a decreasing function. The converse is not true; a one-one mapping does not imply that the function is increasing or decreasing. Cases (i), (ii) and (iv) of the diagram are all one-one mappings, i.e. functions with inverses. Cases (i) and (ii) are increasing functions; case (iv) is not increasing or decreasing. Hence, an inverse function can always be written for an increasing (or decreasing) function; but inverses exist in other cases.

9.4. Limits of sequences. The exploitation of the real number system is effected by introducing and applying the concept of a 'limit'. The concept itself is implicit in the definition of real numbers and we can get, easily enough, a general idea of what it implies.

When we write $\sqrt{2}=1{\cdot}4142\ldots$ or $\pi=3{\cdot}14159\ldots$, we mean that the real number can be approximated by rationals, the more closely the greater number of decimal places we take. The real number itself is, in some sense, the 'limiting value' as the number of decimal places is increased without end. The difficulty (faced in 2.4) is to get a precise formulation, as a foundation to carry the weight of the super-structure of properties contructed on it. The same difficulty appears in designing a precise and general definition of a 'limit'. It is essential to achieve precision here since the whole of the calculus rests on the foundation of limit processes.

The properties of the number system come into play in various ways. Suppose a function $y=f(x)$ is defined for all real x. The set of all real numbers is indefinitely extended in the sense that there are always x's larger than any specified value, and it is indefinitely dense in the sense that there are always x's between any two specified values. If you think of a number, no matter how large, I can always produce a larger one; if you think of two numbers, no matter how close together, I can always produce one between them. We are thus led to ask the questions: what happens to $y=f(x)$ as larger and larger real numbers are assigned to x, and what happens when real numbers are assigned closer and closer to some given real number α? The answer to the first question turns on the concept of the limit of $f(x)$ as x increases without bound, written $\operatorname*{Lim}_{x\to\infty} f(x)$. The other question is answered by defining the limit of $f(x)$ as x approaches α, written $\operatorname*{Lim}_{x\to\alpha} f(x)$. In pursuing these matters, we find at some stage that we require the complete ordered property of real numbers. Suppose $f(x)$ is bounded over some interval of x. Then, because $f(x)$ has an upper bound, it must have a LUB; because $f(x)$ has a lower bound, it must have a GLB. It is because the LUB and GLB exist that we can pin down the limit.

It is convenient to start with a particular case, the limit of a sequence. The basic idea here is that a sequence such as 1·4, 1·41, 1·414, 1·4142, ... has a limit $\sqrt{2}$. The convenience lies in the fact that a sequence is easy to handle, with a simple graphical representation. But the case is important enough in itself, with applications to be pursued in Chapter 11.

Consider the function $f(x)$ defined on some domain of real numbers

x which includes all positive integers 1, 2, 3, ... Write the sequence of real values:

$$f(1), f(2), f(3), \ldots f(n), \ldots .$$

Elementary algebra provides many examples, e.g. sequences (or series) of terms and the corresponding sequences of sums of terms. A well-known case is that of the G.P. (geometric progression): $1, r, r^2, r^3, \ldots r^{n-1}, \ldots$ for any real r. The sum of n terms is:

$$1+r+r^2+\ldots+r^{n-1}=\frac{1-r^n}{1-r}$$

which is itself a sequence:

$$1, \frac{1-r^2}{1-r}, \frac{1-r^3}{1-r}, \ldots \frac{1-r^n}{1-r}, \ldots .$$

To illustrate the possibilities fairly fully, consider five particular cases of sequences, shown graphically in Fig. 9.4 by plotting $f(n)$ against $n=1, 2, 3, \ldots$:

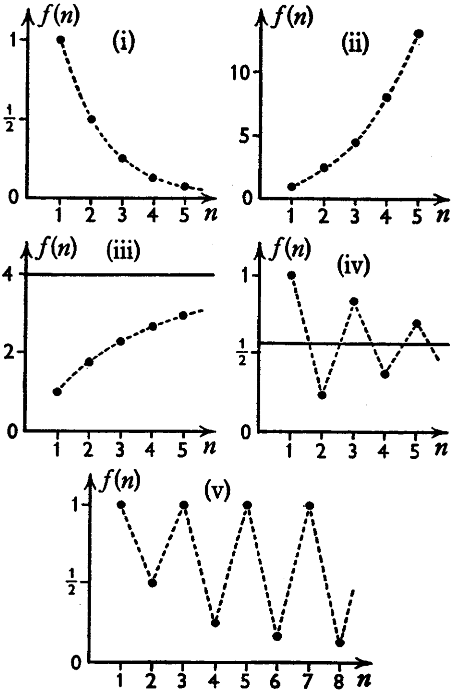


FIG. 9.4

(i) $f(n) = (\frac{1}{2})^{n-1}$.

The first five terms are plotted:

$$1, \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{16}, \dots .$$

These are the terms of a G.P. with $r = \frac{1}{2}$.

(ii) $f(n) = \dfrac{(\frac{3}{2})^n - 1}{\frac{3}{2} - 1} = 2\{(\frac{3}{2})^n - 1\}$

or the sum of n terms of a G.P. with $r = \frac{3}{2}$. Again the first five terms are plotted:

$$1, \tfrac{5}{2}, \tfrac{19}{4}, \tfrac{65}{8}, \tfrac{211}{16}, \dots .$$

(iii) $f(n) = \dfrac{1 - (\frac{3}{4})^n}{1 - \frac{3}{4}} = 4\{1 - (\frac{3}{4})^n\}$

i.e. the sum of n terms of a G.P. with $r = \frac{3}{4}$:

$$1, \tfrac{7}{4}, \tfrac{37}{16}, \tfrac{175}{64}, \tfrac{781}{256}, \dots .$$

(iv) $f(n) = \dfrac{1 - (-\frac{3}{4})^n}{1 + \frac{3}{4}} = \frac{4}{7}\{1 - (-\frac{3}{4})^n\} = \frac{4}{7}\{1 + (\frac{3}{4})^n\}$ (n odd)

and $= \frac{4}{7}\{1 - (\frac{3}{4})^n\}$ (n even)

which is also the sum of a G.P. with $r = -\frac{3}{4}$:

$$1, \tfrac{1}{4}, \tfrac{13}{16}, \tfrac{25}{64}, \tfrac{181}{256}, \dots .$$

(v) $f(n) = 1$ (n odd) and $\dfrac{1}{n}$ (n even).

Here the first eight terms are plotted: $1, \frac{1}{2}, 1, \frac{1}{4}, 1, \frac{1}{6}, 1, \frac{1}{8}, \dots$.

The question is: what happens to $f(n)$ as n increases (through integral values) without bound. To save words, the notation '$n \to \infty$' is used to stand for 'n increases without bound'; it means no more and no less than this. If we wish, we can read '$n \to \infty$' as 'n tends to infinity'; but, if we do, we must remember that '∞' of 'infinity' has no meaning by itself. Certainly, we must avoid any suggestion whatever that n can be put equal to '∞'.

There seems to be little difficulty with any of the five examples. The existence or otherwise of a limit is evident both from the form of $f(n)$ and from the graph. A term like $(\frac{1}{2})^n$ or $(\frac{3}{4})^n$, and more generally r^n for a given positive number $r < 1$, gets smaller and approaches the limit zero. On the other hand, a term r^n where $r > 1$, e.g. $(\frac{3}{2})^n$, gets larger and increases without bound. Equally, terms like $\dfrac{1}{n}$ or $\dfrac{1}{n^2}$ tend to zero, terms in n or n^2 increase without bound. As $n \to \infty$, the conclusions appear:

(i) $f(n) = (\frac{1}{2})^{n-1} \to 0$ steadily, i.e. $\underset{n\to\infty}{\text{Lim}} (\frac{1}{2})^{n-1} = 0$

(ii) $f(n) = 2\{(\frac{3}{2})^n - 1\} \to \infty$, increasing without bound

(iii) $f(n) = 4\{1 - (\frac{3}{4})^n\} \to 4$ steadily, i.e. $\underset{n\to\infty}{\text{Lim}} 4\{1 - (\frac{3}{4})^n\} = 4$

(iv) $f(n) = \frac{4}{7}\{1 - (-\frac{3}{4})^n\} \to \frac{4}{7}$ through oscillations,

i.e. $\underset{n\to\infty}{\text{Lim}} \frac{4}{7}\{1 - (-\frac{3}{4})^n\} = \frac{4}{7}$

(v) $f(n) = 1$ (n odd) and $\dfrac{1}{n}$ (n even) oscillates with no limit.

These are all confirmed and illustrated by the graphs.

Two points need to be stressed. First, we have simplified matters by writing the sequence $f(n)$ as n takes *integral* values 1, 2, 3, ... There is no corresponding simplicity in the values of $f(n)$; these are real numbers, not integers. In the five examples, it happens that the values of $f(n)$ are rationals, but irrationals can easily appear, e.g. if $f(n)$ involves $(\surd 2)^n = 2^{\frac{1}{2}n}$. We must expect all the difficulties associated with the real number system to arise in defining limits. Second, there are clearly several different cases to watch for; some are represented in the examples and there may well be others. For this reason alone, we must proceed with great care if we are not to overlook something. In any case, for such a basic concept, we must make the definition precise and the development systematic.

9.5. The limit process. As is shown formally in 15.4, it is not easy to achieve precision; we have to go down very far into the fundamental ideas of what a limit process is. The immediate object here is to try out these ideas in this particular, and simplified, case of the limit of a sequence.

In considering limits for a function $y = f(x)$, we have to specify two things. The first is a set of *stages* for handling the variation of x. This is easy enough in the present case of the sequence $f(n)$ for $n = 1, 2, 3, \ldots$. The stages are a countably infinite set, i.e. the sequence of integers 1, 2, 3, ..., or better still the sequence of sets:

Stage I = set of integers $p \geqslant 1$; Stage II = set of integers $p \geqslant 2$; ...

and generally:

$$\text{Stage } N = \text{set of integers } p \geqslant n.$$

The successive stages are marked off as the segments I, II, III, ... along *On* in the diagram. The second specification is the *limit process* to which the values of $y = f(x)$ are subject as x varies through its stages. In the present case, consider all the values $y = f(p)$ for stage

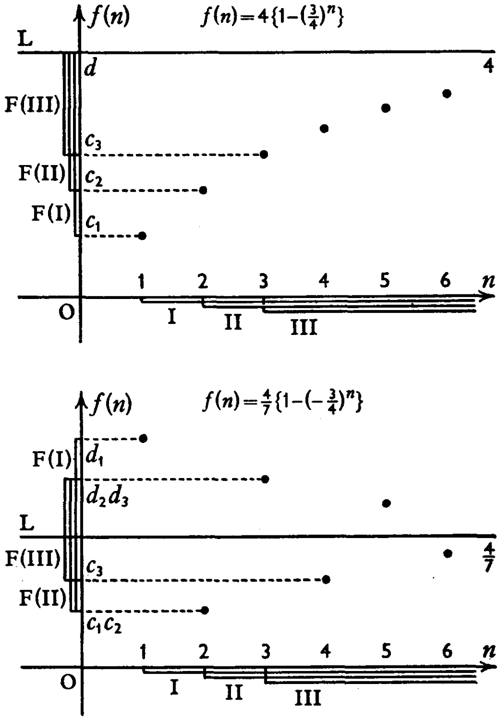


FIG. 9.5

$N(p \geqslant n)$, making up a set Y of real numbers. If the values are not bounded, there is no limit process. If they are bounded, then (by the complete ordered property of real numbers) the set Y has a lower bound and so a GLB c_n, and it has an upper bound and so a LUB d_n. All $y = f(p)$ of Y are contained in the interval $[c_n, d_n]$, i.e. $c_n \leqslant y \leqslant d_n$, and this is the *smallest* such interval.* Denote it by $F(N)$: the least

* Note that the interval $[c_n, d_n]$ comprises *all* real numbers y such that $c_n \leqslant y \leqslant d_n$. Amongst them are the particular real numbers $y = f(p)$ for $p \geqslant n$. Generally there are many other real numbers in the interval.

interval containing all $f(p)$ for stage $N\,(p \geqslant n)$. Hence, to the sequence of stages I, II, III, ... N, ... there corresponds a sequence of intervals F(I), F(II), F(III), ... $F(N)$, ... By the definition, each interval is contained in preceding intervals, i.e. the sequence $F(N)$ forms a shrinking nest of intervals. Fig. 9.5 illustrates two cases. Hence:

DEFINITION: *For $f(n)$ as n increases without bound and for a sequence of stages N (integers $p \geqslant n$), a* **limit process** *exists if $f(n)$ is bounded in each N. It is the nest of decreasing intervals $F(N)$, where $F(N)$ is the smallest interval containing all $f(p)$ over stage $N\,(p \geqslant n)$.*

Consider now the intersection of all intervals $F(N)$, i.e. the set composed of real numbers common to all $F(N)$. Let $F(N) = [c_n, d_n]$ and let $F(M) = [c_m, d_m]$ for a later stage $(m > n)$. Then $F(M)$ is contained in $F(N)$, i.e. $c_m \geqslant c_n$ and $d_m \leqslant d_n$. Hence, as we advance through the stages, c_n increases and, since it has an upper bound (e.g. any d_m), it has a LUB c; similarly, d_n decreases with a GLB d. Hence an interval $[c, d]$ is defined, either one point $(c = d)$ or a finite interval $(c < d)$. Any point common to all $F(N)$ must belong to $[c, d]$. For, if y is one such, then $c_n \leqslant y \leqslant d_n$ all n. So, y is an upper bound of the set of c_n's, i.e. $y \geqslant$LUB c. Similarly, $y \leqslant$GLB d. Hence, $c \leqslant y \leqslant d$ and y belongs to $[c, d]$. This development, which uses the complete ordered property of real numbers, establishes the important and powerful result:

THEOREM: *If the limit process $F(N)$ over stages N exists, then the intersection of all intervals $F(N)$ is itself an interval $F = [c, d]$, where $c \leqslant d$, called the* **final residue** *of the limit process.*

It is now clear that there are only three possibilities, three different things which can happen to $f(n)$ as n increases without bound:

I No limit process exists; the values of $f(n)$ are not bounded.

II A limit process exists and the final residue $F = [c, d]$ is a finite interval $(c \quad d)$; we say that the limit of $f(n)$ does not exist.

III A limit process exists and the final residue F is the single real number L $(c = d = L)$; we say that $f(n)$ has limit L.

DEFINITION: *As n increases without bound, $f(n)$ is* **convergent** *to the* **limit** *L if the limit process $F(N)$ over stages N exists and if the final residue F consists of a single real number L. Write:*

$$\mathit{Lim}_{n\to\infty} f(n) = L \quad \mathit{or} \quad f(n) \to L \quad \mathit{as}\ n \to \infty.$$

The possibilities are illustrated by the examples of 9.4. Example (ii) has $f(p)=2\{(\frac{3}{2})^p-1\}$, not bounded in any stage $N(p\geqslant n)$, i.e. case I. Here, $f(n)$ increases without bound as n increases without bound. This can be written:

$$2\{(\tfrac{3}{2})^n-1\}\to\infty \quad \text{as } n\to\infty.$$

This notation means no more and no less than the statement just made. Case II is illustrated by example (v) where $f(p)=1$ (p odd) and $f(p)=\dfrac{1}{p}$ (p even), i.e. $F(N)$ is the interval $[0, 1]$ at every stage $N(p\geqslant n)$. $\operatorname*{Lim}_{n\to\infty} f(n)$ does not exist. Case III is illustrated by the other examples which give limits:

(i) $\operatorname*{Lim}_{n\to\infty}(\frac{1}{2})^{n-1}=0$; (ii) $\operatorname*{Lim}_{n\to\infty}4\{1-(\frac{3}{4})^n\}=4$; (iii) $\operatorname*{Lim}_{n\to\infty}\frac{4}{7}\{1-(-\frac{3}{4})^n\}=\frac{4}{7}$.

It is, of course, case III which is the one of main interest.

As far as the limit of a sequence $f(n)$ is concerned, we are now through. From the strict definition put forward, we know precisely what a limit means. In particular we have isolated the three possibilities: I $f(n)$ not bounded; II $f(n)$ varying within a finite interval without converging; III $f(n)$ converging to a single real number L. We have a complete set of categories; we know we are overlooking nothing. In practice, we need not take long to reach a conclusion. Given a function $f(n)$, we first see whether or not $f(n)$ is bounded. If $f(n)$ is not bounded, we write $f(n)\to\infty$ as $n\to\infty$ as a matter of notation.* If $f(n)$ is bounded, we check whether or not the intervals containing $f(p)$ for $p\geqslant n$ shrink down to a single number L as n increases. If they do, we write $f(n)\to L$ as $n\to\infty$. We need, here, to keep an eye open for the 'odd case out' where $f(p)$ oscillates in a finite interval for $p\geqslant n$, no matter how large n; this is the case of no limit.

Various necessary and sufficient conditions for a limit can be deduced from the definition given here. One is often used:

THEOREM: *$f(n)$ converges to the limit L as $n\to\infty$ if and only if, for a given positive number ϵ (however small), there is a stage $N(p\geqslant n)$ such that:*

$$L-\epsilon\leqslant f(p)\leqslant L+\epsilon \quad \textit{for all } p\geqslant n.$$

Directly: if $f(n)\to L$ as $n\to\infty$, there must come a stage N when the

* Or we can write $f(n)\to-\infty$ as $n\to\infty$ if $f(n)$ takes unbounded *negative* values.

shrinking interval $F(N)$, comprising $f(p)$ for $p \geqslant n$, is so small around L that it is contained in the given interval $[L-\epsilon, L+\epsilon]$, no matter how small ϵ is. Hence $L-\epsilon \leqslant f(p) \leqslant L+\epsilon$ for $p \geqslant n$. Conversely: if n can be found so that $L-\epsilon \leqslant f(p) \leqslant L+\epsilon$ for given ϵ (however small) and for $p \geqslant n$, then $f(n)$ is bounded and a limit process $F(N)$ exists. As the smallest interval containing $f(p)$ for $p \geqslant n$, $F(N)$ is contained in $[L-\epsilon, L+\epsilon]$. By choosing ϵ small enough (and n large enough to match), $[L-\epsilon, L+\epsilon]$ excludes any specified number $\neq L$, i.e. $F(N)$ can be made to exclude any number $\neq L$. The final residue F can include only L, i.e. $f(n) \rightarrow L$ as $n \rightarrow \infty$. Q.E.D.

This theorem gives conditions which are necessary and sufficient, i.e. which are equivalent to the definition of a limit given here. The conditions are, in fact, those often offered as the definition of a limit, a perfectly valid procedure. However, they are neither as clear conceptually, nor yet as practically useful, as the definition adopted above. It is no easy matter to put up various (small) values of ϵ and then to check for each whether an appropriate (and perhaps very large) integer n does exist. See 9.9, Ex. 21.

9.6. Limits of functions. After this preliminary survey, we turn to the main question: given a function $y=f(x)$ defined on a domain X of real numbers, how do we define the limit of y as x approaches a particular value α? It might be thought that there is no problem here, that the limit is just $f(\alpha)$. This jumps to conclusions too quickly (see 9.7 below). The next idea might be to write a sequence of x's converging on α, and use the concept of a limit of a sequence already developed. But this won't do. The integers are a countable set, a sequence. The real numbers are not countable and a sequence of them does not cover all their properties. We must face this added complication, not avoid it. Specifically, the stages of any limit process must be re-defined to allow for variation of real numbers and not just for a sequence of integers. This is a problem which has bothered mathematicians since Newton and Leibniz developed the calculus in the seventeenth century. The approach adopted here is that of Moore and Smith (1922).*

The real-valued function $y=f(x)$ is considered for real values of

* E. H. Moore and H. L. Smith: 'A General Theory of Limits', *American Journal of Mathematics*, Vol. 44, p. 102.

x around a specified value α and with a view to defining $\lim_{x\to\alpha} f(x)$.*
Strictly we need assume no more than that the domain of x is such that every neighbourhood of α contains some elements of the domain. For convenience, and with little loss of generality, we assume that $f(x)$ is defined on an interval $[a, b]$ of real numbers containing α within it, i.e. for $a \leqslant x \leqslant b$ where $a < \alpha < b$. We allow the exception that $f(x)$ may not be defined at $x = \alpha$. A wider domain such as all x or $x \geqslant a$ would do equally well. A neighbourhood of α is denoted generally as $N = [a_n, b_n]$ where $a \leqslant a_n < \alpha < b_n \leqslant b$. The set S of all possible neighbourhoods N provides the stages for x approaching α through real numbers. S is essentially a non-countable set and any sequence of neighbourhoods (e.g. a contracting nest of intervals) is a very special subset. Among the neighbourhoods in S are some pairs, N_1 and N_2, such that N_2 is contained in N_1. If N_1 is $[a_{n1}, b_{n1}]$ and N_2 $[a_{n2}, b_{n2}]$, then $a_{n1} \leqslant a_{n2} < b_{n2} \leqslant b_{n2}$. If N_1 and N_2 are so related, we write $N_2 > N_1$ and say that N_2 is more advanced than N_1. The relation '$>$' or 'more advanced' simply means 'contained in'. It is a transitive relation, so that, if $N_3 > N_2$ and $N_2 > N_1$, then $N_3 > N_1$. On the other hand, not all neighbourhoods of S are so related. Any two of them, N_1 and N_2, must overlap (not disjoint) since they all contain α; but one need not be contained in the other. The important property of S is that, since N_1 and N_2 overlap, there must be a third neighbourhood N_3 contained in both of them, as in Fig. 9.6a. Formally:

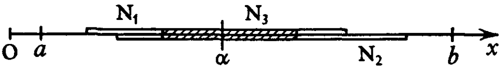


FIG. 9.6a

If N_1 and N_2 are any two neighbourhoods of α, then there exists a third neighbourhood N_3 so that $N_3 > N_1$ and $N_3 > N_2$.

As stages, the neighbourhoods N are such that, whatever pair we pick, we can always find another which is more advanced than either, i.e. contained in both.

A limit process can now be specified very much as before, using the complete ordered property of real numbers. If $y = f(x)$ is not bounded

* As a special case, take the limit of $f(x)$ as x increases without bound ($x \to \infty$). Here we can either extend the idea of the limit of $f(n)$ as $n \to \infty$ or put $x = 1/p$ in $f(x)$ and let $p \to 0$. See 9.9 Ex. 23, 29 and 30.

in a neighbourhood N, then there is no limit process. If it is bounded in every neighbourhood $N=[a_n, b_n]$, then the bounded values assumed by y for x in N are confined in a *smallest* interval $[c_n, d_n]$. Hence, if $a_n \leqslant x \leqslant b_n$, then $c_n \leqslant y \leqslant d_n$. Write $F(N)=[c_n, d_n]$ corresponding to $N=[a_n, b_n]$. From this definition, it follows at once that, if N_2 is more advanced than (contained in) N_1, then the interval $F(N_2)$ is contained in the interval $F(N_1)$. So:

DEFINITION: *For $y=f(x)$ as x approaches α and for stages*

$$N=[a_n, b_n]$$

as neighbourhoods of α, a **limit process** *exists if $f(x)$ is bounded in each N. It is the set of smallest intervals $F(N)=[c_n, d_n]$ such that $c_n \leqslant y \leqslant d_n$ when $a_n \leqslant x \leqslant b_n$ and with the property:*

$$\textit{if } N_2 > N_1 \textit{ then } F(N_2) \textit{ is contained in } F(N_1).$$

Fig. 9.6*b* illustrates for a simple function $y=x(x+1)$ defined on $x>0$ and considered around $x=2$, $y=6$. Three neighbourhoods N_1, N_2 and N_3 are shown, with $N_3>N_1$ and $N_3>N_2$. The corresponding $F(N_1)$, $F(N_2)$ and $F(N_3)$ have the property that $F(N_3)$ is contained in $F(N_1)$ and $F(N_2)$.

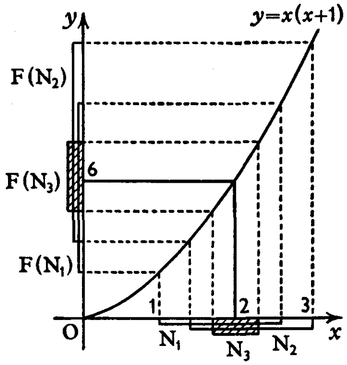


FIG. 9.6*b*

The essential result, very simple and yet very powerful, is obtained as before, by use of the complete ordered property of real numbers:

THEOREM: *If the limit process $F(N)$ over stages N exists, then the intersection of all intervals $F(N)$ is itself an interval $F=[c, d]$, where $c \leqslant d$, called the* **final residue** *of the limit process.*

Proof: *any* two neighbourhoods N_1 and N_2 must overlap and contain a third, and *any* two intervals $F(N_1)$ and $F(N_2)$ must overlap and contain a third. The lower end of one interval $F(N_1)$ can be above the upper end of another $F(N_2)$ only if these intervals are disjoint, and this is ruled out. If $F(N)=[c_n, d_n]$, then all c_n's (as N varies) are less than or equal to all d_n's. Since the whole set of c_n's is bounded above, there is a LUB c. Since the whole set of d_n's is bounded below, there is a GLB d, and $c \leqslant d$. Hence an interval $F=[c, d]$ is defined, either a single point ($c=d$) or a finite interval ($c<d$). It remains to show that all y common to all $F(N)$ are in F. Take any such y: $c_n \leqslant y \leqslant d_n$ for all $N=[a_n, b_n]$. Hence y is an upper bound of the set of all c_n's and so $y \geqslant$LUB c. Similarly, $y \leqslant$GLB d. Hence $c \leqslant y \leqslant d$ and y belongs to $F=[c, d]$. Q.E.D.

The definition of a limit then follows:

DEFINITION: *As x approaches α, $f(x)$ is* **convergent** *to the* **limit** *L if the limit process $F(N)$ over stages N exists and if the final residue F consists of a single real number L. Write:*

$$\mathit{Lim}_{x \to \alpha} f(x)=L \quad \textit{or} \quad f(x) \to L \quad \textit{as } x \to \alpha.$$

As a consequence of the definition, necessary and sufficient conditions for a limit can be laid down as follows:

THEOREM: *$f(x)$ converges to the limit L as $x \to \alpha$ if and only if, for a given positive number ϵ (however small), there is a stage $N=[a_n, b_n]$ such that:*

$$L-\epsilon \leqslant f(x) \leqslant L+\epsilon \quad \textit{for all } x \textit{ of } N\,(a_n \leqslant x \leqslant b_n).$$

The proof is similar to that of the corresponding result of 9.5 above. Since these are necessary and sufficient conditions, i.e. equivalent to the definition of a limit, it is perfectly in order to use them alternatively as the definition.

There are just three possibilities to consider and to distinguish carefully:

I No limit process exists for $f(x)$ at $x=\alpha$. The values of $f(x)$ are not bounded in neighbourhoods of $x=\alpha$.

II A limit process exists for $f(x)$ at $x=\alpha$ and the final residue $F=[c, d]$ is a finite interval ($c<d$) so that $\operatorname{Lim}_{x \to \alpha} f(x)$ does not exist.

III A limit process exists for $f(x)$ at $x=\alpha$ and the final residue F is the single real number L so that $\operatorname{Lim}_{x \to \alpha} f(x)$ exists and equals L.

Several examples illustrate:

(i) $y=1/1-x$ defined on all $x\,(x \neq 1)$. No limit process exists at $x=1$ since neighbourhoods (excluding $x=1$ itself where y is not defined) are such that y is not bounded. Here y increases without bound (numerically, through positive or negative values) as x approaches 1. We write: $1/1-x \to \pm\infty$ as $x \to 1$, but this notation means no more than the statement given.

(ii) $y =$ least integer not less than x defined on $x > 0$. This is the step-function graphed in Fig. 9.6c (as in 8.6). A limit process exists at $x = 1$. Write $N = [a_n, b_n]$ where $0 < a_n < 1$ and $b_n > 1$. Then N gives the interval of y's: $F(N) = [1, 4]$ if $3 < b_n \leqslant 4$; $F(N) = [1, 3]$ if $2 < b_n \leqslant 3$; $F(N) = [1, 2]$ if $1 < b_n \leqslant 2$. Hence the final residue $F = [1, 2]$. There is no limit of y as $x \to 1$.

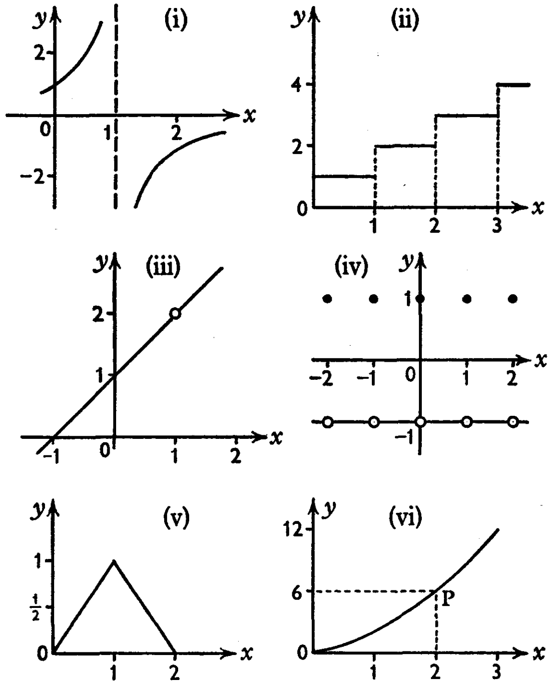


FIG. 9.6c

(iii) $y = \dfrac{1-x^2}{1-x}$ defined on all $x\,(x \neq 1)$. If $x \neq 1$,

$$y = \frac{(1-x)(1+x)}{1-x} = 1+x\,.$$

If $x = 1$, y is not defined. Hence the function is the same as the linear function $1+x$, except that the point $x = 1$ must be omitted. The graph (Fig. 9.6c) is the line $y = 1+x$ with a gap at $x = 1$. A limit process exists at $x=1$ and neighbourhoods N (excluding always $x=1$ itself) are such that the intervals $F(N)$ always cover $y=2$ and contract onto this value. Hence $\operatorname*{Lim}_{x\to 1} \dfrac{1-x^2}{1-x} = 2$. More generally:

$$\operatorname*{Lim}_{x\to \alpha} \frac{\alpha^n - x^n}{\alpha - x} = n\alpha^{n-1},$$

as can be seen by dividing $\alpha - x$ into $\alpha^n - x^n$.

(iv) $y=1$ (x integral) and $y=-1$ (x non-integral) defined on all x. The graph is reproduced in Fig. 9.6*c* exactly as in 9.2 above. Let $N=[a_n, b_n]$ be a neighbourhood of $x=1$ (excluding $x=1$ itself). Then the interval of y's: $F(N)=[-1, 1]$ if $a_n \leqslant 0$ or $b_n \geqslant 2$; $F(N)$ is the single number -1 if $0<a_n<1$ and $1<b_n<2$. Hence the limit process exists with final residue $F=-1$, i.e.

$$\operatorname*{Lim}_{x\to 1} y = -1 \quad \text{whereas } y=1 \text{ at } x=1.$$

Contrast with the function:

$$y=1\ (x \text{ rational}) \quad \text{and} \quad y=-1\ (x \text{ irrational}).$$

The graph *looks* like two lines $y=1$ and $y=-1$ but, in fact, if there is a point on one line, there is no point on the other at a given x. Here, for all neighbourhoods N of $x=1$, the interval of y's is $F(N)=[-1, 1]$. There is no limit.

(v) $y=1-\surd(x-1)^2$ defined on $0 \leqslant x \leqslant 2$. It is easily seen that a limit process exists at $x=1$ and that the final residue is 1. Hence $\operatorname*{Lim}_{x\to 1} \{1-\surd(x-1)^2\}=1$. Notice that $y=1$ at $x=1$ also.

(vi) $y=x(x+1)$ defined on $x>0$. This is the usual kind of well-behaved function, the one used above to illustrate the limit process. For $x=2$, $y=6$ and $\operatorname*{Lim}_{x\to 2} y=6$, at P on the curve in Fig. 9.6*c*. A similar result holds for other values of x, at other points on the curve.

Of the three possibilities, I is illustrated by example (i), II by (ii) and III by the considerable range of cases of (iii)–(vi).

9.7. Continuity. Some of the cases where a limit exists are of more interest, and more useful, than others. In particular, the last two cases (v) and (vi) of 9.6 can be separated off from the others, by the property that the limit of $f(x)$ exists at $x=\alpha$ *and* that the limit is the same as $f(\alpha)$. It is this property which serves to describe what we mean by a 'continuous' function. So:

DEFINITION: *$f(x)$ is* **continuous** *at $x=\alpha$ if* (1) *$\underset{x\to a}{Lim} f(x)$ exists,* (2) *$f(\alpha)$ is defined and* (3) *$\underset{x\to a}{Lim} f(x)=f(\alpha)$.*

Continuity is a property of a function at a particular value, of a curve at a particular point. It represents the agreement between limit and value where both exist. As examples, (v) and (vi) are continuous at $x=1$ and $x=2$ respectively. On the other hand, (i)–(iv) are all discontinuous at $x=1$, (i) because neither the function nor the limit is defined there, (ii) because there is no limit, (iii) because the function is not defined and (iv) because the limit and value (both existing) are different. As a natural extension, $f(x)$ is *continuous over the domain X* if it is continuous for each x of X. Then (v) and (vi) are continuous everywhere; (i) and (iii) are continuous except at $x=1$; (ii) and (iv) have each a countably infinite number of discontinuities.

It must be noted that continuity is essentially a characteristic of the real number system. It is pointless to ask whether $f(n)$, a function of integral n, is continuous or not. Hence a variable such as n (integral values) is called a *discrete variable* while x taking all real numbers (e.g. in an interval) is a *continuous variable.* The term *continuum* is sometimes applied to the set of all real numbers, to the set of all points on a directed line.

In conclusion, we return to the question of whether $\underset{x\to a}{\text{Lim}} f(x)$ can be defined or found by means of a *sequence* of intervals or values of x converging on α. If the limit L is known to exist, *any* sequence of neighbourhoods of α (or of particular values within them) which shrink down to α must define a sequence of intervals (or values) of $f(x)$ which converges to L. This is *necessarily* so. But it is not *sufficient* to establish that the limit exists. A sequence is not enough when a strict development of the limit concept is attempted — or in practice when the existence of a limit is in question. It may well do in practice, however, when we are sure that the limit exists. The following calculations illustrate.

Consider $f(x)=x(x+1)$ at $x=2$. Write $x=2-h$ and $x=2+k$ to get:

$$f(2-h)=(2-h)(3-h)=6-5h+h^2;$$
$$f(2+k)=(2+k)(3+k)=6+5k+k^2.$$

It is easily established that the neighbourhood $[2-h, 2+k]$ of $x=2$ gives $f(x)$ in the interval $[6-5h+h^2, 6+5k+k^2]$. The limit process for $x\to 2$ is now expressed as letting h and k take, quite separately, smaller and smaller positive real values. The limit is seen to exist and to be 6. Now short-circuit the process by taking intervals $[2-h, 2+h]$ and by specifying only a sequence of h's, by writing $h=\frac{1}{n}$, where $n=1, 2, 3, \ldots$. So:

$$f\left(2-\frac{1}{n}\right)=6-\frac{5}{n}+\frac{1}{n^2} \quad ; \quad f\left(2+\frac{1}{n}\right) = 6+\frac{5}{n}+\frac{1}{n^2} \,.$$

Both tend to 6 as $n\to\infty$. Hence, if the limit of $f(x)$ as $x\to 2$ is known to exist, then it is 6. But this sequential process does not establish that the limit exists. It is easy to produce a case where no limit exists but where a sequential process suggests one. Consider the function $y=1$ (x rational) and $y=-1$ (x irrational) at, say, $x=2$. Then $f\left(2-\frac{1}{n}\right)=f\left(2+\frac{1}{n}\right)=1$ (all n) since $2\pm\frac{1}{n}$ are rational. Both have limit 1 as $n\to\infty$, suggesting that $f(x)\to 1$ as $x\to 2$. There is, in fact, no such limit.

9.8. Properties of limits and continuity. If limits exist for several functions $f(x)$, $g(x)$, ... at $x=\alpha$, it can be shown that the limit of a combination of the functions is the combination of the separate limits. Writing Lim for $\lim\limits_{x\to\alpha}$:

$$\begin{aligned}
&\text{Lim}\,\{f(x)+g(x)\}=\text{Lim}\,f(x)+\text{Lim}\,g(x);\\
&\text{Lim}\,\{f(x)-g(x)\}=\text{Lim}\,f(x)-\text{Lim}\,g(x);\\
&\text{Lim}\,\{f(x)\times g(x)\}=\text{Lim}\,f(x)\times\text{Lim}\,g(x);\\
&\text{Lim}\,\frac{f(x)}{g(x)}=\frac{\text{Lim}\,f(x)}{\text{Lim}\,g(x)} \quad \text{for Lim}\,g(x)\neq 0.
\end{aligned}$$

Here it must be checked that the domains of both $f(x)$ and $g(x)$ are appropriate to the definition of limits as $x\to\alpha$, e.g. that $f(x)$ and $g(x)$ are defined on an interval containing α. Further, for a composite function:

$$\text{Lim}\,F\{f(x)\}=F\{\text{Lim}\,f(x)\}$$

provided that $F(u)$ is defined on an interval containing $u=L$, where $L=\text{Lim}\,f(x)$.

The proofs of these results follow from the definition of a limit but they need to be laid out formally. One proof, that for the sum result, is as follows. Suppose that $f(x)$ and $g(x)$ are defined on some interval containing α, $f(x)$ converging to L and $g(x)$ to L' as $x \to \alpha$. The same set of stages N can be used for each: $N=[a_n, b_n]$ as a neighbourhood of α on which $f(x)$ and $g(x)$ are defined. To N there corresponds a smallest interval $F(N)=[c_n, d_n]$ containing $f(x)$ for x in N, and a smallest interval $G(N)=[c_n', d_n']$ containing $g(x)$ for x in N. There is also a smallest interval $H(N)$ for values of $h(x)=f(x)+g(x)$ for x in N. Then $H(N)$ is contained in the interval $[c_n+c_n', d_n+d_n']$ formed from $F(N)$ and $G(N)$. But $F(N)$ has final residue L and $G(N)$ final residue L', so that $[c_n+c_n', d_n+d_n']$ converges to the single value $L+L'$. Hence, there is for the function $h(x)$ a limit process $H(N)$ over stages N and the final residue is $L+L'$. Lim $h(x)$ exists and it is $L+L'$, i.e. Lim $\{f(x)+g(x)\}=$ Lim $f(x)+$ Lim $g(x)$. Q.E.D.

There are corresponding results for continuity. If $f(x)$ and $g(x)$ are continuous at $x=\alpha$, then $f(x)+g(x)$, $f(x)-g(x)$, $f(x)\times g(x)$ and $\dfrac{f(x)}{g(x)}$ are all continuous at $x=\alpha$. In the last case, $g(\alpha)\neq 0$ must be assumed. Further, if $y=F(u)$ is continuous at $u=f(\alpha)$ and if $u=f(x)$ is continuous at $x=\alpha$, then the composite function $y=F\{f(x)\}$ is continuous at $x=\alpha$. Similar results hold for functions which are continuous over a domain X, i.e. continuous at each x of X.

These results follow from those for limits and only the last causes any trouble. Write $L=f(\alpha)$ and $M=F(L)=F\{f(\alpha)\}$. Then we job backwards as follows. Using the conditions for a limit (theorem of 9.6), we assign an interval $[M-\epsilon, M+\epsilon]$ for values of y, where we take ϵ as small as we please. Since $y=F(u)$ is continuous at $u=L$, corresponding to $y=M$, there is an interval (neighbourhood) N' of u around L such that the corresponding values of $y=F(u)$ are contained in $[M-\epsilon, M+\epsilon]$. Since $u=f(x)$ is continuous at $x=\alpha$, corresponding to $u=L$, there is an interval (neighbourhood) N of x around α such that the values of $u=f(x)$ are contained in N'. Fig. 9.8*a* illustrates. Hence, no matter how small ϵ, there is a neighbourhood N of α so that $M-\epsilon \leqslant F\{f(x)\} \leqslant M+\epsilon$ for all x of N. By the theorem of 9.6 again, $F\{f(x)\}$ converges to $M=F\{f(\alpha)\}$ as $x \to \alpha$, i.e. $F\{f(x)\}$ is continuous at $x=\alpha$. Q.E.D.

In dealing with continuous functions, e.g. in problems of maxima and minima, we make constant use of the following result. The property is an important one but very simple; it is, indeed, so simple that we tend to accept it without question. However, it is something which needs to be established; since the proof is somewhat tricky and tedious, it is left to 15.4 below. The result is:

THEOREM: *If $y=f(x)$ is continuous over the interval $[a, b]$, then the range of y is itself an interval $[c, d]$.*

It is essential to appreciate the import of this property of continuous functions. It is not only that the values of y are encompassed by an interval $[c, d]$, but also that y ranges over the whole interval. For any x in $[a, b]$, the corresponding value of y is in $[c, d]$. Conversely, if y is any value in $[c, d]$, then there is some x in $[a, b]$ which corresponds. Fig. 9.8*b* illustrates. Two particular cases are important in themselves. They refer to the attainment of smallest and largest values, and to the consequences of a change in sign, of $f\ (x)$ over $[a, b]$. See 9.9 Exs. 31 and 32.

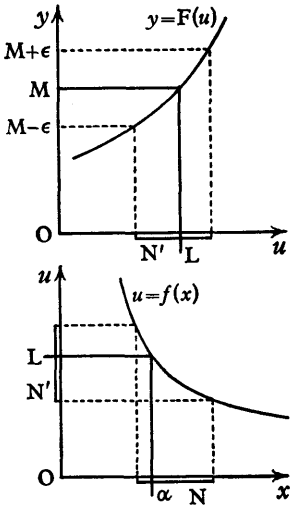


FIG. 9.8*a*

Fig. 9.8*b* indicates that particular attention should be paid to functions which are both continuous and increasing (or decreasing). If $y=f(x)$ is an increasing function defined on the domain A and with range B, then the inverse $y=f^{-1}(x)$ exists and it is an increasing function defined on the domain B and with range A. The sets A and B need not be intervals and the functions $f(x)$ and $f^{-1}(x)$ need not be continuous (see Fig. 9.3). As a special case, we have the following useful result, for an increasing function as in (iii) of Fig. 9.8*b*:

THEOREM: *If $y=f(x)$ is increasing and continuous over the interval $[a, b]$, then*

(1) *the range of $f(x)$ is the interval $[c, d]$, where $c=f(a)$ and $d=f(b)$;*

(2) *the inverse $y=f^{-1}(x)$ exists, an increasing and continuous function over the interval $[c, d]$, with the interval $[a, b]$ as range.*

This follows from the preceding theorem. A similar result, with obvious variations as in (iv) of Fig. 9.8*b*, holds for a decreasing and continuous function.

To conclude with some words of encouragement: it is necessary to have a strict definition of a limit and hence of continuity of a function, both to serve as a sound theoretical basis for the calculus and also as a practical guide to the range of possibilities which can occur. However, in practice, the 'odd case out' is usually clear enough, as at the end of 9.7. Further, the writing of limits of *simple* functions is easy enough in practice. More complicated functions need to be split into combinations of simple functions; the results above then apply. An illustration is given in 9.9 Ex. 29. Hence the final point: the *practice* of limit evaluation is not at all difficult.

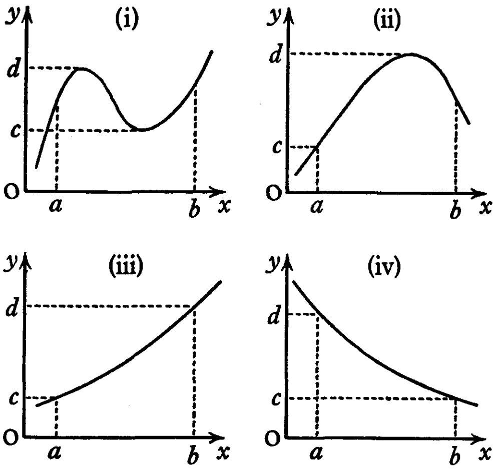


FIG. 9.8*b*

9.9. Exercises

1. Show that $y = 1 + x^2$ can be defined for all x with range $y \geqslant 1$. Represent graphically as a locus.

2. *Functions in parametric form.* Illustrate the use of parameters by showing that the general ratio of quadratics $y = \dfrac{ax^2 + bx + c}{\alpha x^2 + \beta x + \gamma}$ includes both (i) and (ii) of 9.2, and that (iii) of 9.2 can be included in $y = \dfrac{ax + b + c\sqrt{x - \alpha}}{\sqrt{x - \beta}}$.

3. Extend the notion of parametric classing of functions by showing that a more general step-function than (v) of 9.2 is $y=\lambda_r$ for $ra \leqslant x < (r+1)a$ $(r=0, 1, 2, \ldots)$.

4. *The parabola.* Show that $y=x^2$, defined for all x, is a two-one mapping with range $y \geqslant 0$. Limit the domain to $x \geqslant 0$ and show that the mapping is one-one. Write the inverse function and show that both the function and inverse are increasing. Generalise to $y=ax^2$ (parameter $a \neq 0$). See 8.9 Ex. 21 and 22.

5. Show that $y_1=\dfrac{(x+a)^2-a^2}{x}$ and $y_2=x+2a$ take the same values for all x except at $x=0$ where $y_2=2a$ but y_1 is not defined.

6. *Functions as mappings.* Two ways of mapping a function (7.3 and 8.6) are mentioned in 9.2. Link by showing that $y=f(x)$ as a curve effects a mapping of points on one line Ox onto points on another line Oy (at right angles) as in Fig. 9.9*a*.

7. Exhibit function (ii) of 9.2 as a composite function $F\{f(x)\}$ where $f(x)=x^2$. Put function (iv) of 9.2 into the same form. Specify the domains and ranges concerned.

8. By writing as functions of functions and checking domains and ranges, show that $y=\sqrt[3]{1-\sqrt{1+x}}$ is defined for $x \geqslant -1$ but that $y=\sqrt{1-\sqrt{1+x}}$ only for $-1 \leqslant x \leqslant 0$.

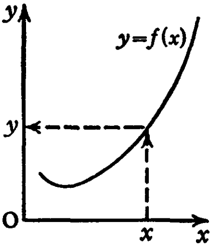


FIG. 9.9*a*

9. Illustrate a practical use of functions by considering the following. A firm makes open tin cans in the form of a cylinder of height 1 foot and of variable radius x feet. The cost of production is £y per 1000 cans where $y=\sqrt{u}$ and where u is the surface area of the can in square feet. Show that, as a function of a function, $y=\sqrt{\pi x(x+2)}$ for $x>0$.

10. If x and y each take all values in the interval $[-1, 1]$, the relation $x^2+y^2=1$ is shown by a circle. Show that no function is defined on this domain. Restrict the set from which x or y is drawn in various ways to get y as a function of x or inversely, indicating which segments of the circle you use. If the domain of x must be an interval, show that a one-one relation (and so a function and inverse) arises only if the interval is some sub-interval either of $[-1, 0]$ or of $[0, 1]$.

11. Show that the linear function $y=mx+c$ $(m \neq 0)$ is monotonic (either increasing or decreasing) and hence can be inverted.

12. Show that $y=\sqrt{x^2}-1$ is defined for all x as a two-one relation, but that it is a decreasing function on $x \leqslant 0$ and an increasing function on $x \geqslant 0$.

13. Graph $y=\dfrac{1+x^2}{1-x^2}$ defined on $x>1$. Show that the inverse is $x=\sqrt{\dfrac{y-1}{y+1}}$ defined on $y<-1$, both functions being increasing.

14. Show that the function $y=\frac{1}{2}x$ $(0<x \leqslant 1)$, $=\frac{1}{2}(x-1)$ $(2<x \leqslant 4)$ is in-

creasing, but that $y=\frac{1}{2}x$ $(0<x\leqslant 1)$, $=\frac{1}{2}x-2$ $(2<x\leqslant 4)$ is not. On the other hand, show that both have inverses, as (ii) and (iv) of Fig. 9.3.

15. *Strictly increasing functions.* The definition of 9.3 is that of a strictly increasing function, excluding all step-functions. Weaken to: $f(x)$ is increasing if $a<b$ implies $f(a)\leqslant f(b)$ and show that step-functions are included.

16. Show that $\dfrac{1}{2+n}\to 0$ and $\dfrac{1}{2-n}\to 0$ as $n\to\infty$. Represent graphically and show that the tendency is steady (through positive and negative values respectively), provided that $n>2$ in the second case. Show further that, as $n\to\infty$:

$$\frac{2+n}{1+n}\to 1;\ \frac{2-n}{1+n}\to -1;\ \frac{2+n}{1-n}\to -1;\ \frac{2-n}{1-n}\to 1\ .$$

17. Generalise Ex. 16 by taking

$$f(x)=(a_r x^r+a_{r-1}x^{r-1}+\dots)/(\alpha_s x^s+\alpha_{s-1}x^{s-1}+\dots)$$

where $a_r\neq 0$, $\alpha_s\neq 0$. As $n\to\infty$, show that $f(n)\to 0$ if $r<s$, that $f(n)\to\dfrac{a_r}{\alpha_s}$ if $r=s$ and that there is no limit process, $f(n)\to\pm\infty$, if $r>s$.

18. Illustrate the case of a limit process but no limit by showing that $f(n)=(\frac{1}{2})^n+(-1)^n$ tends to oscillate between ± 1 as $n\to\infty$.

19. *Geometric series.* Establish formally that $a+ar+ar^2+\dots+ar^{n-1}$ has sum $S_n=a\dfrac{1-r^n}{1-r}\to\dfrac{a}{1-r}$ if $|r|<1$, and that $S_n\to\pm\infty$ if $|r|>1$, as $n\to\infty$. What can be said of the cases where $r=\pm 1$?

20. *Arithmetic progression.* Show that there is no limit process for the sum of n terms of $a+(a+\alpha)+(a+2\alpha)+\dots$: $S_n=\frac{1}{2}n\{2a+(n-1)\alpha\}\to\pm\infty$ as $n\to\infty$.

21. If $f(n)=\frac{4}{7}\{1-(-\frac{3}{4})^n\}$, show that $\frac{4}{7}-\epsilon\leqslant f(n)\leqslant\frac{4}{7}+\epsilon$ for all $n\geqslant 7$ given $\epsilon=0{\cdot}1$ but that we must take $n\geqslant 15$ if $\epsilon=0{\cdot}01$ is given. On the other hand, if $f(n)=1/n^2$, the convergence (to zero) is more rapid. Show that, given ϵ, then $n\geqslant 1/\sqrt{\epsilon}$ suffices for $0\leqslant f(n)\leqslant\epsilon$, e.g. $\epsilon=0{\cdot}1$, $n\geqslant 4$; $\epsilon=0{\cdot}01$, $n\geqslant 10$.

*22. *Area of a circle.* The Euclidean definition of the area of a circle is the limit of the area of a regular inscribed polygon of n sides (as $n\to\infty$). Hence, for a circle of radius r and area A: $A=\lim\limits_{n\to\infty}\frac{1}{2}nr^2\sin\dfrac{360^\circ}{n}$ (see 2.9 Ex. 8). It is still not established that the limit exists, still less what it is. But, given $\dfrac{\sin x}{x}\to 1$ as $x\to 0$ (x in radians, $180^\circ=\pi$ radians), show that $(n/2\pi)\sin 2\pi/n\to 1$ as $n\to\infty$ and so $A=\pi r^2$. Archimedes (*circa* 250 B.C.) worked out π as between 223/71 and 22/7 for $n=96$; he was probably the only mathematician before Newton and Leibniz in the seventeenth century with any real idea of a limit.

23. Show that the limit process for $f(n)$ as n increases without bound through integers (9.5) extends to the limit process of 9.6 for $f(x)$ as x increases without bound through real values and that the same three possibilities arise:

no limit process or $f(x)\to\pm\infty$; no limit; $f(x)\to L$ as $x\to\infty$.

24. As $x \to -1$, find whether there are limits of (i) $y=1+x$; (ii) $y=1-x$; (iii) $y=(1-x^2)/(1+x)$; (iv) $y=1/(1+x)$; (v) $y=1/(1-x)$, and (vi) $y=(1+x)/(1-x^2)$. What is the difference between the functions (ii) and (iii), and between (v) and (vi)?

25. Show that the function $y=-1$ $(x<-1)$; $y=x$ $(-1 \leqslant x \leqslant 1)$; $y=1$ $(x>1)$ has a graph as in Fig. 9.9*b*. Examine from the point of view of limits and continuity at $x=\pm 1$.

26. For the functions of Ex. 5, show that y_1 and y_2 both have limit $2a$ as $x \to 0$. Why is y_1 (unlike y_2) discontinuous at $x=0$?

27. Show that the function $y=x$ $(0 \leqslant x \leqslant 1)$; $y=x-1$ $(2<x \leqslant 3)$ has inverse $x=y$ $(0 \leqslant y \leqslant 1)$; $x=y+1$ $(1<y \leqslant 2)$. Represent graphically. Why are these functions (though defined) not continuous at $x=1$ and $y=1$ respectively?

28. If $f(x)$ is continuous over $[a, b]$, can we say that $y=f(x)$ has inverse $x=f^{-1}(y)$ over $[a, b]$ if and only if $f(x)$ is monotonic? See Fig. 9.3.

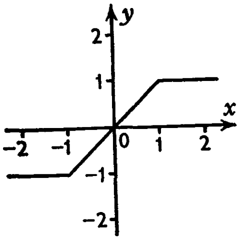


Fig. 9.9*b*

29. Express $f(x)=x/\surd(1+x^2)$ as a function of u where $u=1+1/x^2 \to 1$ as $x \to \infty$. Deduce that $f(x) \to 1$ as $x \to \infty$.

30. Alternatively, put $x=1/h$ and show that

$$x/\surd(1+x^2)=1/\surd(h^2+1) \to 1 \text{ as } h \to 0 \text{ so that } x/\surd(1+x^2) \to 1 \text{ as } x \to \infty$$

31. *Weierstrass' Theorem.* Establish from the theorem of 9.8 that:

If $f(x)$ is continuous over $[a, b]$ then $f(x)$ attains its smallest value Inf $f(x)$ at some α $(a \leqslant \alpha \leqslant b)$ and similarly for its largest value Sup $f(x)$. Illustrate by reference to Fig. 9.8*b* and check graphically that $y=x^2-2x-1$ over $[0, 3]$ has its smallest value at $x=1$ and its largest at $x=3$.

32. *Bolzano's Theorem.* Establish further that:

If $f(x)$ is continuous over $[a, b]$ and such that $f(a)<0$ and $f(b)>0$ then $f(\alpha)=0$ for some α $(a<\alpha<b)$.

Illustrate graphically. Check that $x^2-2x-1=0$ has a root between 2 and 3.

***33.** *The field of convergent sequences of rationals.* A sequence of *rational* values $a_1, a_2, \ldots a_n, \ldots$ is convergent if $a_n \to a$ as $n \to \infty$, where a is some *real* value. Use the theorem of 9.5 to express the condition for convergence: the sequence is convergent if and only if, given a positive rational ϵ (however small), there is an integer n such that $|a_p - a_q| \geqslant \epsilon$ for all integers p and $q \geqslant n$. So convergent sequences of rationals can be handled without reference to their limits, i.e. without reference to real numbers. In particular, they can be shown to be a field. This is the basis of Cantor's definition of a real number as a convergent sequence of rationals, an alternative to Dedekind's definition (2.4 and 15.1).

CHAPTER 10

CALCULUS

10.1. Some examples. A few simple examples serve to indicate both the nature of the main concepts of the calculus and their great range of applications.

(i) *Speed, velocity and rate of growth.* A car is driven steadily at 60 m.p.h., travelling 88 feet every second. What is the distance x feet covered in t seconds? The answer is easy: $x=88t$. Hence, in this particular case of constant speed, we have the following simple functions of time:

$$\text{Distance covered at time } t\text{: } x=88t\text{; Speed at time } t\text{: } v=88.$$

Here t is in seconds (from some starting time), x in feet, v in feet per second.

Suppose, now, that the car is driven at a varying speed. Then there arise such questions as: what is the speed when $\frac{1}{4}$ mile has been travelled? Or, how long does it take the car to accelerate from rest to 60 m.p.h.? Whatever the speed may be, suppose observations are made of the distance x feet travelled for various elapsed times of t seconds and suppose they disclose that $x=2t^2$, distance as a particular function of time. Select some elapsed time t_0 seconds and perform the simple calculations:

Distance travelled in 10 seconds from t_0

$$=2(t_0+10)^2-2t_0^2=2t_0^2+40t_0+200-2t_0^2=(40t_0+200)\text{ feet.}$$

Average speed over 10 seconds from $t_0=(40t_0+200)/10=(4t_0+20)$ feet per second. Other such calculations can be made and summarised, p. 265. There is a practical difficulty which we have glossed over. We have said that observations 'disclose' that $x=2t^2$, precisely; surely they do no such thing? Actual readings may be limited to multiples of 1/10th second and 1/10th foot. Then, strictly, we can do no more than get to the third row of the table below. However, a simple fact

Time interval	*Average speed u*	
	ft. per sec.	*over time of*
$[t_0, t_0+10]$	$u=4t_0+20$	10 secs.
$[t_0, t_0+1]$	$u=4t_0+2$	1 sec.
$[t_0, t_0+0{\cdot}1]$	$u=4t_0+0{\cdot}2$	0·1 sec.
...	...	...
$[t_0, t_0+h]$	$u=4t_0+2h$	h secs.

stands out: the limiting speed or velocity* is $4t_0$. We want to take up the position that time t varies continuously, that distance x varies continuously, that they are related by $x=2t^2$ precisely, and that velocity v is given by: $v=4t$. Such a step from a practical situation to a theoretical one is one commonly taken but it must be recognised for what it is. In practice, we measure time (say) to 1/10th second; conceptually, we have no qualms in assuming time varies continuously. The same thing is true of distance or velocity. There is then no residual difficulty in assuming a precise formula: $x=2t^2$ or $v=4t$. Hence, provided that we make the appropriate assumptions, we take the big step forward and write $\lim\limits_{h\to 0} u=v$. We call v the instantaneous speed or velocity at the specified elapsed time. So, at elapsed time t_0:

$$\text{Average speed over } [t_0, t_0+h]=4t_0+2h;\ \text{Velocity at } t_0=4t_0.$$

The first depends on t_0 and h, the second on t_0 only. Hence, at time t:

$$\text{Distance travelled } x=2t^2;\ \text{Velocity } v=4t$$

dropping the subscript 0. There are two functions of time t (in seconds elapsed), one for x (in feet) and the other for v (in feet per second). The second function is derived from the first; it is an example of a 'derivative'.

The question we put can now be answered. After time t, the car has accelerated to $v=60$ m.p.h. $=88$ feet per second. What is t? Answer: $4t=v=88$, i.e. $t=22$. The distance travelled is then

$$x=2(22)^2=968.$$

* Strictly, velocity is a vector quantity with magnitudes and direction, where 'magnitude' corresponds to the length of a geometric vector (8.4). Here we speak only of magnitude, e.g. for movement along a straight path.

Hence, the car reaches 60 m.p.h. after travelling 968 feet in 22 seconds. This is on the basis of the observations from which the formula $x = 2t^2$ is obtained.

The functions can be plotted on two graphs (Fig. 10.1*a*). The first curve shows how the distance x travelled increases with time t; after elapsed time $ON = t$, the distance is $NP = 2t^2$. The second curve (a line) shows how velocity v increases with time t; after elapsed time $ON = t$, the velocity is $NQ = 4t$. The second curve is here derived from the first. Can we reverse the process, i.e. given the velocity in terms of time, can we find the distance travelled? We have, at least, a hint of how to obtain this 'anti-derivative', of how to perform this 'integration'. Given the point Q on the graph of the velocity ($v = 4t$ at t), calculate:

Area of triangle ONQ

$$= \tfrac{1}{2}ON \times NQ = \tfrac{1}{2}t \times 4t = 2t^2$$

which we recognise as the height of the graph of distance at the same t. Hence, in this case, the *height* NP of the distance curve ($x = 2t^2$ at t) is obtained as the *area* under the velocity line from O to t. This also fits in with the dimensions of the concepts involved: distance travelled is speed × time. That is: we expect the height of the distance graph to be related to areas on the velocity graph. This is the fascinating relationship which will concern us later on.

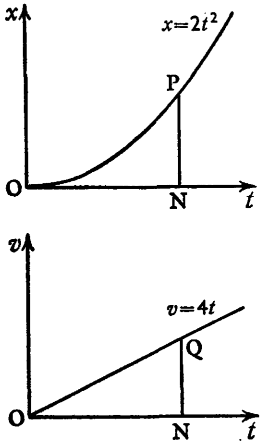


FIG. 10.1*a*

Other such examples can be quoted. To take one more, suppose x is the number (thousands) unemployed after t months from some starting date and suppose observations disclose that $x = 2t^2$. There is the same problem of the step from practice to theory, since the numbers unemployed may only be recorded monthly, for $t = 1, 2, 3, \ldots$. There is, again, no conceptual difficulty in assuming that unemployment varies continuously over time. Making this assumption and performing the same calculations as before, we get the average rate of growth of unemployment over h months from month t as $4t + 2h$, and hence the (instantaneous) rate of growth at month t as $4t$. Hence, after t months:

Unemployed $x=2t^2$ (thousands);
Rate of growth $v=4t$ (thousands per month).

For example, after 10 months, the number unemployed is 200,000, growing at the rate of 40,000 per month. The rate of growth is the 'derivative' of the number; the number is the 'anti-derivative' or 'integral' of the rate of growth.

(ii) *Average and marginal rate of change.* A firm produces gadgets at a given average (unit) cost of £20 each in a plant with a capacity of 100 gadgets per week. Total cost is £$20x$ ($0 \leqslant x \leqslant 100$) for a weekly output of x gadgets. This is the particular case of fixed average cost. On the other hand, suppose that total cost £y varies with output according to the formula: $y=2x^2$. What can be said about the cost of getting additional output? If weekly output is running at x_0 gadgets, then the average (unit) cost of h more gadgets per week is obtained as before. It is u of the table above, with t_0 replaced by x_0, with seconds replaced by gadgets per week, and with u in terms of £ per gadget. As $h \to 0$, $u \to v$ and v is called the marginal cost, i.e. the marginal rate of change of total cost:

Average cost over $[x_0, x_0+h]=4x_0+2h$; Marginal cost at $x_0=4x_0$.

From the given function (total cost $y=2x^2$) is derived a second function: marginal cost $v=4x$. Here y in £ and v in £ per gadget are functions of the weekly output x gadgets. Plotting as graphs as above, we obtain the total cost curve ($y=2x^2$) and the marginal cost curve ($v=4x$); the second is the 'derivative' of the first and the first is got by 'integration' from the second.

Further such examples can be got from other fields. Suppose a roller is pushed up a lawn of increasing slope. Then it may be found that the force v lbs. against the roller increases with the distance x feet travelled, say by the formula $v=4x$. The work done y foot-lbs. also increases with the distance travelled: $y=2x^2$. The relationship here is the same as that just considered. The force is the marginal rate of change of work done. From the work done graph ($y=2x^2$) is 'derived' the graph of the force operating ($v=4x$) and, conversely, the graph of the force can be 'integrated' to give the work done.

(iii) *Slope of chord and tangent to a curve.* We have considered a function $y=2x^2$ and we have derived a second function $v=4x$. If x is time, we interpret v as the instantaneous rate of growth of y over time; if x is some physical variable (e.g. distance or output), we interpret v as the marginal rate of change of y with respect to x. Now concentrate attention on the curve which represents the function $y=2x^2$ in the plane Oxy, as shown in Fig. 10.1*b*.

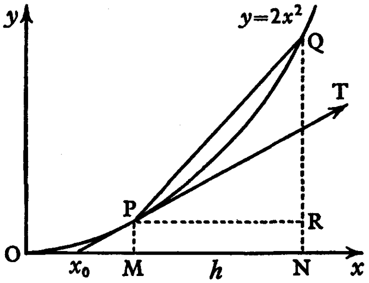


FIG. 10.1*b*

Let P be the point on the curve at x_0 and Q at x_0+h so that $OM=x_0$, $MN=h$. Then the chord PQ has slope:

$$\frac{RQ}{PR}=\frac{NQ-NR}{MN}$$
$$=\frac{NQ-MP}{ON-OM}$$
$$=\frac{2(x_0+h)^2-2x_0^2}{(x_0+h)-x_0}$$

i.e. Slope of $PQ=4x_0+2h$ (average rate of change).

As $h\to 0$, Q approaches P and the chord PQ tends to a limiting position, the tangent PT at P. So:

Slope of $PT=4x_0$ (marginal rate of change).

To summarise, from a function such as $y=2x^2$, we have derived a second function $v=4x$. Here v is the marginal rate of change of y with respect to x or (if x happens to be time) the instantaneous rate of growth of y over time. Further, v measures the slope of the tangent to the curve $y=2x^2$ at the appropriate point P. A study of v, as the 'derivative' of y, will clearly be rewarding in terms of the range of its applications.

10.2. Derivatives. The calculus is based on the assumption that a function $y=f(x)$ is a continuous real-valued function of a continuous real variable. The domain may be all x or some 'half' set such as $x>0$, but at least it should include some interval $[a, b]$ where $a<b$. We take x as varying through all real numbers in an interval $(a\leqslant x\leqslant b)$ and y as varying continuously over the interval. As an exception, we may allow the function to have one or more isolated discontinuities, e.g. $y=1/(1-x)$ which is discontinuous (and not defined) at $x=1$. The first concept of the calculus is the 'derivative', arising out of the considerations of 10.1:

DEFINITION: *The function $f(x)$ has* **derivative** *at $x=x_0$:*

$$f'(x_0)=\underset{h\to 0}{Lim}\frac{f(x_0+h)-f(x_0)}{h}$$

if the limit exists and the **derived function** *$f'(x)$ is defined on the domain of real values of x for which derivatives exist.*

The following remarks are needed in amplification of the definition. If x_0 is given, within the domain of x, then the expression

$$\frac{f(x_0+h)-f(x_0)}{h}$$

is the *average rate of change* of $f(x)$ over the interval $[x_0, x_0+h]$ of varying length h. But h must be regarded as taking both positive and negative real numerical values; the interval, in fact, is *either* $[x_0, x_0+h]$ if $h>0$, *or* $[x_0-(-h), x_0]$ if $h<0$. On the other hand, the expression is not defined for $h=0$. Hence it is a function of h, not defined for $h=0$, but defined at all other points in a neighbourhood of $h=0$. If the expression has a limit as $h\to 0$ through all real values, i.e. through *both* positive *and* negative real h, then we get the derivative $f'(x_0)$ as the limit of the average rate of change of $f(x)$ from $x=x_0$. This is the *marginal rate of change* of $f(x)$ at $x=x_0$. (If there is no limit, there is no derivative and no marginal rate of change.) As a definition of rates of change and an interpretation of a derivative:

DEFINITION: *If $f(x)$ has derivative $f'(x_0)$, the* **marginal rate of change** *of $f(x)$ at $x=x_0$ is the limit (as $h\to 0$) of the* **average rate of change** $\frac{f(x_0+h)-f(x_0)}{h}$ *and it is measured by $f'(x_0)$.*

When there is no ambiguity, $f'(x)$ can be called the *rate of change* of $f(x)$.

The essential point about a derivative is that it *assumes* continuous variation, leaving a gap between theory and practice. In practice, x takes discrete values, e.g. time in 1/10ths seconds, which can be recorded. Only the average rate of change of $f(x)$ is relevant. This may be all we need. But, usually, we assume that x is capable of continuous variation and that y varies continuously with x. We can then write the marginal rate of change of $f(x)$, the derivative $f'(x)$. This is generally accepted procedure, e.g. for distance varying continuously with time as in (i), or for cost varying continuously with output as in (ii) of 10.1. The fact that the assumption is made should not be overlooked.

Various notations for a derivative of $y=f(x)$ are in common use and it is convenient to have them. Each can be applied to y or to $f(x)$:

(i) y' or $f'(x)$ following Lagrange (1736–1813)
(ii) $D_x y$ or $D_x f(x)$ following Cauchy (1789–1857)
(iii) $\frac{d}{dx}y$ or $\frac{d}{dx}f(x)$ following Leibniz (1646–1716).

Of these, (i) is quite appropriate when the function is unspecified, while (ii) is particularly useful for a particular function, e.g.

$$D_x(2x^2)=4x.$$

The symbol D_x, or more simply D when there is no ambiguity, is to be regarded as an operator. Thus $D(2x^2)=4x$ means that $4x$ is the function obtained by operating on $2x^2$ by writing its derivative at each x. To indicate the derivative, the result of the operator D, at a particular value x_0, we can write $[D(2x^2)]_{x_0}=4x_0$. The oldest of the notations is (iii) and it can be quite safely used if it is understood that '$\frac{d}{dx}$' is a single symbol, the operator $D=\frac{d}{dx}$. However, the notation used to be employed, and still is sometimes, in the form $\frac{dy}{dx}$, described as a 'differential coefficient'; the danger here is that $\frac{dy}{dx}$ gets separated into the ratio of 'dy' to 'dx', which is without meaning (at least in the present context).

The following is an important result:

THEOREM: *A necessary condition that $f'(x_0)$ exists is that $f(x)$ is continuous at $x=x_0$, i.e. if $f'(x_0)$ exists, then $f(x)$ is continuous at $x=x_0$.*

The proof involves the properties of limits (9.8). If $f'(x_0)$ exists, write $x=x_0+h$ and let $x \to x_0$ $(h\to 0)$: $\frac{f(x)-f(x_0)}{x-x_0}\to f'(x_0)$ as $x\to x_0$.

Write: $$F(x)=\frac{f(x)-f(x_0)}{x-x_0}(x-x_0)+f(x_0).$$

Hence $F(x)=f(x)$, except that $F(x)$ is not defined, whereas $f(x)$ is, at $x=x_0$. Take the limit of $F(x)$ as $x\to x_0$ (which does not involve any value at $x=x_0$):

$$F(x)\to f'(x_0)\times 0+f(x_0)=f(x_0).$$

Hence $f(x)\to f(x_0)$ also, i.e. $f(x)$ is continuous at $x=x_0$. Q.E.D.

It must be stressed that the condition is *necessary*, i.e. whenever $f'(x)$ exists, $f(x)$ is continuous. It is *not sufficient*, i.e. if $f(x)$ is continuous, then $f'(x)$ may or may not exist. To illustrate, consider

$$\begin{aligned} f(x) = 1 - \surd\{(1-x)^2\} &= 1-(1-x) = x \qquad (x \leqslant 1) \\ &= 1-(x-1) = 2-x\,(x>1) \end{aligned}$$

Then for variation around $x=1$:

$$\frac{f(1+h)-f(1)}{h} = 1 \quad (h<0) \quad \text{and} \ -1 \quad (h>0).$$

There is no limit as $h \to 0$ through positive and negative values (10.9 Ex. 3). The function $f(x)$ is continuous but without derivative at $x=1$.

The graphical version of a derivative makes the position clear. If P and Q are the points $x=x_0$ and $x=x_0+h$ respectively on the curve $y=f(x)$, then:

$$\text{Slope of } PQ = \frac{f(x_0+h)-f(x_0)}{h} \quad (h<0 \text{ or } h>0)$$

as in 10.1, (iii) above. Hence:

DEFINITION: *The* **tangent** *PT to the curve $y=f(x)$, at P where $x=x_0$, is the line through P with slope* $\underset{h\to 0}{Lim}\dfrac{f(x_0+h)-f(x_0)}{h}=f'(x_0)$, *if the derivative exists. Otherwise there is no tangent at P.*

The necessary condition is that, if a tangent exists at P, then the curve is continuous at P. The condition is not sufficient, i.e. if the curve is continuous at P, then there may or may not be a tangent at P. The case of failure occurs when the curve has a sharp point. In Fig. 10.2, the function (defined for all $x>0$) is discontinuous at $x=1$ and continuous but without derivative at $x=2$. The curve jumps at the point P_1 of discontinuity; it does not jump, but the tangent does, at the sharp point P_2. To the left of P_2, the tangent slopes upwards; to the right, it slopes downwards.

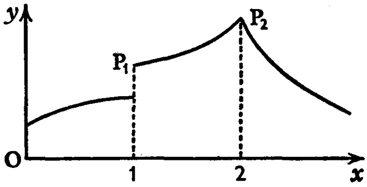


FIG. 10.2

Hence, even when $f(x)$ is continuous, it is necessary to ensure that $f'(x)$ exists before it is written. The next result is for a composite function:

THEOREM: *If $f(x)$ has derivative $f'(x_0)$ and if $F(u)$ has derivative $F'(u_0)$ at $u_0=f(x_0)$, then the composite function $F\{f(x)\}$ has derivative at x_0:*

$$[F'(u)]_{u_0=f(x_0)} \times f'(x_0)$$

i.e. $D_xF\{f(x)\}=D_uF(u)\times D_xf(x)$ *where* $u=f(x)$.

Proof: Write $u_0=f(x_0)$ and $u_0+k=f(x_0+h)$ so that $\frac{k}{h}=\frac{f(x_0+h)-f(x_0)}{h}$.

Then:

$$\frac{F\{f(x_0+h)\}-F\{f(x_0)\}}{h}=\frac{F(u_0+k)-F(u_0)}{k}\,\frac{f(x_0+h)-f(x_0)}{h}.$$

From the properties of limits and since $k\to 0$ as $h\to 0$:

$$\lim_{h\to 0}\frac{F\{f(x_0+h)\}-F\{f(x_0)\}}{h}=\lim_{k\to 0}\frac{F(u_0+k)-F(u_0)}{k}$$
$$\times\lim_{h\to 0}\frac{f(x_0+h)-f(x_0)}{h}$$

i.e. $D_xF\{f(x)\}=D_uF(u)\times D_xf(x)$ at $x=x_0$. Q.E.D.

COROLLARY: *If $y=f(x)$ is continuous and increasing with derivative $D_xf(x)$ at x_0, so that $x=f^{-1}(y)$ exists, is continuous and increasing, then:*

$$D_yf^{-1}(y)=\frac{1}{D_xf(x)} \quad \text{at } y_0=f(x_0) \quad \textit{provided } D_xf(x)\neq 0.$$

Proof: we have $x=f^{-1}(y)=f^{-1}\{f(x)\}$ i.e. $D_x(x)=D_xf^{-1}\{f(x)\}$. By the theorem, $D_xf^{-1}\{f(x)\}=D_yf^{-1}(y)\times D_xf(x)$. Also $D_x(x)=1$ (see 10.3 below). Hence $D_yf^{-1}(y)\times D_xf(x)=1$, which proves the corollary. Clearly it holds equally if $f(x)$ is a decreasing function.

10.3. Operational rules for derivatives. Two situations occur in handling derivatives. The derivative may be needed for a function which is unspecified but written as a combination of other functions. Here we seek *rules* for the derivative of the function in terms of the derivatives of the separate components. In the other situation, the derivative is required for a particular function, e.g. a specified algebraic function. The problem is: given a particular function, write another particular function, the second being the derived function of the first. Here we need to have ready to hand a list of the derivatives, called *standard forms,* of the simplest particular functions; we rely upon the rules to give derivatives of more complicated functions.

As a preliminary, two very special derivatives can be written:

$$D(\text{constant}) = 0 \quad \text{and} \quad D(x) = 1.$$

These follow from the definition and they are evident from the geometric point of view. The line $y = \text{constant}$, parallel to Ox, is its own tangent and it has zero slope. The line $y = x$, also its own tangent, has unit slope.

The first task is to establish, from the definition, rules for the derivatives of combinations of functions. The obvious rules to obtain are those which deal with sums, differences, products and quotients. The proofs of these rules are all similar and sufficiently illustrated by the product rule: If $f(x)$ and $g(x)$ are two functions with derivatives, write $F(x) = f(x)g(x)$ so that

$$\begin{aligned}\frac{F(x+h) - F(x)}{h} &= \frac{f(x+h)g(x+h) - f(x)g(x)}{h}\\ &= f(x)\,\frac{g(x+h) - g(x)}{h} + g(x+h)\,\frac{f(x+h) - f(x)}{h}\\ &\to f(x)g'(x) + g(x)f'(x) \quad \text{as } h \to 0\end{aligned}$$

i.e. $F'(x)$ exists and equals $f(x)g'(x) + g(x)f'(x)$.

The rules are assembled in the following table, together with those for composite and inverse functions (last theorem and corollary, 10.2 above).

Rule	*Function*	*Derivative*
Sum	$f(x) + g(x)$	$f'(x) + g'(x)$
Difference	$f(x) - g(x)$	$f'(x) - g'(x)$
Product	$f(x)g(x)$	$f(x)g'(x) + g(x)f'(x)$
Quotient	$\dfrac{f(x)}{g(x)}$	$\dfrac{g(x)f'(x) - f(x)g'(x)}{\{g(x)\}^2}$ if $g(x) \neq 0$
Composite	$F\{f(x)\}$	$F'(u)f'(x)$ where $u = f(x)$
Inverse	$f^{-1}(x)$	$\dfrac{1}{f'(y)}$ if $f'(y) \neq 0$

All the derivatives written are assumed to exist. In the last case, $f(y)$ is assumed to be increasing (or decreasing) with derivative $f'(y)$, so that the inverse $f^{-1}(x)$ is also increasing (or decreasing) with the derivative shown.

Some particular cases arise from the fact that the derivative of a constant is zero. If A and k are constants, $f(x)+A$ has derivative $f'(x)$ and $kf(x)$ has derivative $kf'(x)$. Write $f(x)=1$, $f'(x)=0$ in the quotient rule to give the derivative of a reciprocal:

$$\frac{1}{g(x)} \quad \text{has derivative} \quad -\frac{g'(x)}{g(x)^2} \quad \text{if } g(x)\neq 0.$$

The rules are of such practical use that it is worth while repeating them in an alternative notation. The particular cases are included. Here u and v are two functions which possess derivatives; in the composite rule, y is a function of u and u is a function of x, both with derivatives.

Rule	*Derivative*	
Additive constant	$D(u+A)=Du$	(A constant)
Sum	$D(u+v)=Du+Dv$	
Difference	$D(u-v)=Du-Dv$	
Multiplicative constant	$D(ku)=kDu$	(k constant)
Product	$D(uv)=uDv+vDu$	
Reciprocal	$D\frac{1}{v}=-\frac{Dv}{v^2}$	if $v\neq 0$
Quotient	$D\frac{u}{v}=\frac{vDu-uDv}{v^2}$	if $v\neq 0$
Composite	$D_xy=D_uyD_xu$	
Inverse	$D_xy=\frac{1}{D_yx}$	if $D_yx\neq 0$

The second task is to establish, from the definition, the derivatives of the simplest functions. At the moment, for algebraic functions, the only simple function to consider is x^r, where r is a rational exponent. Indeed, we can get by with the two special derivatives: $D(\text{constant})=0$ and $D(x)=1$. As quite simple exercises in the application of the rules, the steps are as follows (10.9 Ex. 5 and 6). First, by the product rule and mathematical induction from $D(x)=1$, we obtain $D(x^n)=nx^{n-1}$, n a positive integer. Next, by the particular case of the quotient rule: $D\left(\frac{1}{x}\right)=-\frac{1}{x^2}$. Then, by the composite function rule, with $u=\frac{1}{x}$, we find: $D(x^{-n})=D\left(\frac{1}{x^n}\right)=-\frac{n}{x^{n+1}}=-nx^{-(n+1)}$. Further, write $y=\sqrt{x}$ so that $x=y^2$ with derivative $D_yx=2y$; the inverse function rule gives $D(\sqrt{x})=\frac{1}{2\sqrt{x}}$, i.e. $D(x^{\frac{1}{2}})=\frac{1}{2}x^{-\frac{1}{2}}$. The derivatives of other surds, in general $x^{p/q} = (\sqrt[q]{x})^p$, are obtained similarly. All this boils down to one *standard form*:

$$D(x^r) = rx^{r-1} \quad (r \text{ rational}).$$

With the rules, this is enough for the derivative of any algebraic function.

The fact that $D(\text{constant}) = 0$ is important in another connection. The problem so far is: given $f(x)$, find $f'(x)$. Can the inverse problem also be solved? Given $f(x)$, what function $F(x)$ is such that $F'(x) = f(x)$? If such a $F(x)$ can be found, it can be called an *anti-derivative* of $f(x)$. Clearly, in some cases, but only in some cases, the rules and standard forms supply the anti-derivative, simply by operating them in reverse. So:

$$\text{If } f(x) = rx^{r-1}, \text{ then } F(x) = x^r \text{ has } F'(x) = f(x)$$

i.e.

$$\text{if } f(x) = x^r, \text{ then } F(x) = \frac{x^{r+1}}{r+1} \text{ has } F'(x) = f(x).$$

An anti-derivative of x^r is $\dfrac{x^{r+1}}{r+1}$. Again, suppose that $f(x)$ can be arranged in the form $\phi(x)\psi'(x) + \phi(x)\psi'(x)$, where $\phi(x)$ and $\psi(x)$ are two particular functions, then $F(x) = \phi(x)\psi(x)$ is an anti-derivative sought. For example:

$$f(x) = \frac{1+3x}{2\sqrt{x}} = \frac{1+x}{2\sqrt{x}} + \sqrt{x} = (1+x)D(\sqrt{x}) + \sqrt{x}D(1+x).$$

Hence $F(x) = (1+x)\sqrt{x}$ has derivative $F'(x) = f(x)$ and an anti-derivative of $\dfrac{1+3x}{2\sqrt{x}}$ is $(1+x)\sqrt{x}$.

One question has been forgotten here: if $F(x)$ can be found so that $F'(x) = f(x)$, is $F(x)$ unique? The answer is: not quite. Suppose that $G(x)$ is also such that $G'(x) = f(x)$. Then the derivative of $G(x) - F(x)$ is $G'(x) - F'(x) = f(x) - f(x) = 0$. The only function with a zero derivative everywhere is a constant; the only curve with a tangent everywhere parallel to Ox is a line parallel to Ox. Hence $G(x) - F(x) =$ constant, and the *general* function with derivative $f(x)$ is:

$$F(x) + A \quad \text{where } F'(x) = f(x) \text{ and } A = \text{arbitrary constant}.$$

In other words, if $F(x)$ is any anti-derivative of $f(x)$ which can be found, then the general anti-derivative is obtained by adding any constant to $F(x)$. This lack of uniqueness is important, as we shall see. It expresses the fact that, while multiplicative constants remain, additive constants disappear in derivation.

10.4. Areas. The basic notion of an area is the product of two variables; a rectangle has area $x \times y$ where x and y are the lengths of adjacent sides. In elementary geometry, this leads to the area of a parallelogram (base × height) and of a triangle ($\frac{1}{2}$ base × height), and so to the area of any closed figure bounded by lines. Something quite different is involved when an area is bounded by curves. The elementary approach is to *approximate* the area by inscribing some figure with lines as sides and by writing the area of the inscribed figure. The implicit assumption here is that the curvilinear area is the *limit* of the inscribed area as the fit gets closer. This is the idea behind the definition of the area of a circle in terms of inscribed (or circumscribed) polygons (see 2.9 Ex. 8 and 9.9 Ex. 22).

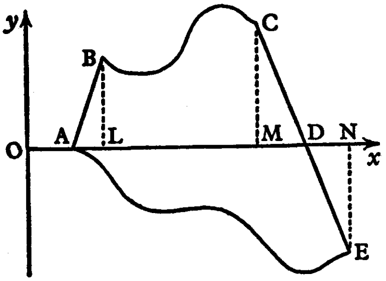


FIG. 10.4*a*

To make this more systematic, we start by assuming, provisionally, that a given figure does indeed have an area. This seems to be in order if the figure is bounded by continuous curves, or by segments of continuous curves and lines which are joined continuously. Then we proceed by two stages. At the first stage, we insert a pair of coordinate axes and divide up the area investigated in the way shown in Fig. 10.4*a*:

$$\text{Area} = ABL + LBCM + MCD + NEA - NDE.$$

Apart from triangles, the component areas are of the type $LBCM$, the area between a curve, the axis Ox and two vertical lines (parallel to Oy). The second stage is to obtain a measure of such an area, given the equation $y=f(x)$ of the curve.

As an actual case, consider the curve of Fig. 10.4*b*, for the function $y=x^2$ defined on the interval [1, 2], and seek a measure of the area $MPQN$ under the curve, above Ox and between MP (at $x=1$) and NQ (at $x=2$). Divide MN into four equal segments, each of length $\frac{1}{4}$, by inserting M_1, M_2 and M_3, with corresponding P_1, P_2 and P_3 on the curve. Complete the four rectangles shown, each being under the curve, so that the area A sought is greater than the sum of the areas of the four rectangles:

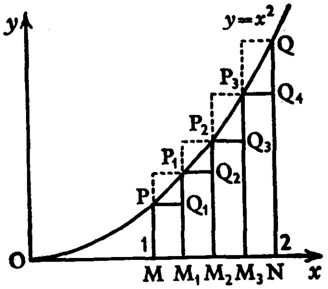


FIG. 10.4b

$$A > MPQ_1M_1 + M_1P_1Q_2M_2 + M_2P_2Q_3M_3 + M_3P_3Q_4N$$
$$= \tfrac{1}{4} \times MP + \tfrac{1}{4} \times M_1P_1 + \tfrac{1}{4} \times M_2P_2 + \tfrac{1}{4} \times M_3P_3$$
$$= \tfrac{1}{4}\{1^2 + (\tfrac{5}{4})^2 + (\tfrac{3}{2})^2 + (\tfrac{7}{4})^2\}$$
$$= \tfrac{1}{64}(4^2 + 5^2 + 6^2 + 7^2) = \tfrac{63}{32}.$$

Similarly, A is less than the sum of the four larger rectangles shown:

$$A < \tfrac{1}{4}\{(\tfrac{5}{4})^2 + (\tfrac{3}{2})^2 + (\tfrac{7}{4})^2 + 2^2\} = \tfrac{1}{64}(5^2 + 6^2 + 7^2 + 8^2) = \tfrac{87}{32}.$$

Hence from this particular *partition* of MN: $\tfrac{63}{32} < A < \tfrac{87}{32}$.

We are not very close, but we can get closer by taking a finer partition of MN, i.e. by approximating by means of larger numbers of thinner rectangles.

Consider, therefore, a partition of MN into n segments each of length $h = \dfrac{1}{n}$. The dividing points are:

$$1,\ 1+h,\ 1+2h,\ 1+3h,\ \dots\ 1+(n-1)h,\ 1+nh=2$$

and the corresponding heights of the curve are the squares of these values. Then, by adding the areas of the n rectangles under the curve:

$$A > h\{1^2 + (1+h)^2 + (1+2h)^2 + \dots + (1+\overline{n-1}h)^2\}$$
$$= h^3\left\{\left(\frac{1}{h}\right)^2 + \left(\frac{1}{h}+1\right)^2 + \left(\frac{1}{h}+2\right)^2 + \dots + \left(\frac{1}{h}+n-1\right)^2\right\}$$
$$= \frac{1}{n^3}\left\{n^2 + (n+1)^2 + (n+2)^2 + \dots + (2n-1)^2\right\}.$$

A simple algebraic result (proved by induction) is that

$$1^2 + 2^2 + 3^2 + \dots + n^2 = \tfrac{1}{6}n(n+1)(2n+1).$$

So:

$$1^2 + 2^2 + 3^2 + \dots + (2n-1)^2 = \tfrac{1}{6}(2n-1)(2n)(4n-1) = \tfrac{1}{3}n(2n-1)(4n-1)$$
$$1^2 + 2^2 + 3^2 + \dots + (n-1)^2 = \tfrac{1}{6}(n-1)(n)(2n-1) = \tfrac{1}{6}n(n-1)(2n-1)$$

i.e. $n^2+(n+1)^2+(n+2)^2+\dots+(2n-1)^2$

$$=\tfrac{1}{3}n(2n-1)(4n-1)-\tfrac{1}{6}n(n-1)(2n-1)$$
$$=\tfrac{1}{6}n(2n-1)(7n-1).$$

Hence: $$A>\frac{n(2n-1)(7n-1)}{6n^3}=\frac{7}{3}\left(1-\frac{1}{2n}\right)\left(1-\frac{1}{7n}\right).$$

In the same way, by adding the areas of the n rectangles above the curve:

$$A<\frac{7}{3}\left(1+\frac{1}{2n}\right)\left(1+\frac{1}{7n}\right).$$

With the n-fold partition, A is found to be contained in the interval:

$$\left[\frac{7}{3}\left(1-\frac{1}{2n}\right)\left(1-\frac{1}{7n}\right),\ \frac{7}{3}\left(1+\frac{1}{2n}\right)\left(1+\frac{1}{7n}\right)\right].$$

In retrospect, what we have achieved is a *definition* of a limit process for the area A, the stages being the sequence of finer partitions as $n=1, 2, 3, \dots$ increases. As $n\rightarrow\infty$, i.e. as the rectangles get thinner without bound ($h\rightarrow 0$), the interval for A converges to 7/3. This is our measure of the area $MPQN$: $A=7/3$.

10.5. Integrals. Given a function $y=f(x)$, defined on an interval $[a, b]$, where $a<b$, the integral from a to b is to be defined in such a way that it is interpreted graphically as the area between the curve $y=f(x)$ and the axis Ox, and between the two lines parallel to the axis Oy at $x=a$ and $x=b$ respectively. If $f(x)$ is never negative on $[a, b]$, the area in question is that shown as A in Fig. 10.5*a*. The development of 10.4, which is essentially algebraic and which involves a limit process, provides the basis for the definition and evaluation of the integral and area.

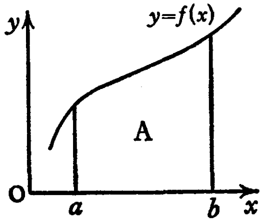


FIG. 10.5*a*

We do need, however, to take a closer look at the development. The limit process used is a sequence of partitions; the interval is split into n equal segments, where $n=1, 2, 3, \dots$ increases without bound. This is not quite good enough for a function $f(x)$ of a continuous variable x. If the integral or area does exist (as assumed at the beginning of 10.4), we can find it by any convenient limit process, e.g. the sequential process adopted. But, to *define* an integral and to *establish* that it exists, we must be careful (as in 9.6) to consider any limit process and not just a particular sequential process. A strict definition, not dependent on a sequence, is given in 15.5. It proceeds on the following lines.

It is assumed only that $y=f(x)$ is bounded on the interval $[a, b]$. A *partition* p of the interval is a set of dividing points $x_0, x_1, x_2, \dots x_n$ where $a=x_0$ and $x_n=b$; otherwise the points can take any values whatever, arranged in ascending order from a to b. The interval is divided into segments of lengths: $x_1-x_0, x_2-x_1, x_3-x_2, \dots x_n-x_{n-1}$. These are not limited, either in number (n) or in lengths, except that they add to $b-a$. Consider the rth segment ($r=1, 2, 3, \dots n$) of length x_r-x_{r-1}. Let x_r' be any point of this segment ($x_{r-1} \leqslant x_r' \leqslant x_r$) and write $f(x_r')$ with x_r' varying over the segment. Then $f(x_r')$, being bounded, must have a GLB L_r and a LUB G_r in the segment. Hence:

$$L_r(x_r-x_{r-1}) \leqslant f(x_r')(x_r-x_{r-1}) \leqslant G_r(x_r-x_{r-1}) \quad \dots\dots\dots\dots(1)$$

The significance of (1) can be seen in terms of Fig. 10.5*b*, which is an enlarged picture of the rth segment. The product $f(x_r')(x_r-x_{r-1})$ is the area of the rectangle shown solid in the diagram, the height being fixed by the value $f(x_r')$ at the selected point x_r'. As x_r' varies over the segment, this product varies with the changing height of the rectangle. Its lower bound is $L_r(x_r-x_{r-1})$, shown by the area of the lower rectangle in Fig. 10.5*b*. Its upper bound is $G_r(x_r-x_{r-1})$, similarly shown as the area of the upper rectangle in the figure. The 'area' under the curve, if it can be defined, is also somewhere in this range.

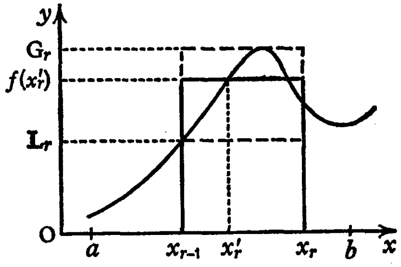


FIG. 10.5*b*

Select a value x_r' in every segment and specify the pair of bounds L_r and G_r for $f(x_r')$ in each case. In many cases, L_r and/or G_r will coincide with one or both of $f(x_{r-1})$ and $f(x_r)$, but this is by no means necessary; the case illustrated above has $L_r=f(x_{r-1})$ and $G_r \neq f(x_r)$. Add the products (1) for all segments, $r=1, 2, 3, \dots n$, to give:

$$\sum_{r=1}^{n} L_r(x_r-x_{r-1}) \leqslant \sum_{r=1}^{n} f(x_r')(x_r-x_{r-1}) \leqslant \sum_{r=1}^{n} G_r(x_r-x_{r-1}) \dots\dots\dots(2)$$

in terms of the $\sum$ notation for sums. The middle sum in (2),

$$\sum_{r=1}^{n} f(x_r')(x_r - x_{r-1}),$$

depends on, and varies with, the selection of the x_r's in the various segments of a given partition P. The other sums, involving L_r and G_r as fixed for each segment, depend in no way on this selection; they depend only on what partition P is given. Write these sums as $L(P)$ and $G(P)$ respectively, to indicate this fact. Hence the sum $\sum_{r=1}^{n} f(x_r')(x_r - x_{r-1})$ is bounded within the given interval for a given P:

$$F(P) = [L(P), G(P)].$$

This interval is the smallest of all such intervals which can be specified.

We now have a limit process. The stages are given by all the various partitions P which can be written. They do not form a sequence. On the other hand, we can say what we mean by advancing through stages: $P_2 > P_1$ can be taken as implying that the partition P_2 is a refinement of P_1 in the sense that P_2 contains all the dividing points of P_1 and some others as well. The value under consideration is the sum $\sum_{r=1}^{n} f(x_r')(x_r - x_{r-1})$, i.e. the sum of all rectangular areas of the kind shown in Fig. 10.5*b*. This is contained in the smallest interval $F(P)$, given P. It is easily seen that, if $P_2 > P_1$, then $F(P_2)$ is an interval contained in $F(P_1)$. The question is: in this limit process, as we advance through finer and finer partitions P, does the interval $F(P)$ converge to a single value or not? If it does, then $\sum_{r=1}^{n} f(x_r')(x_r - x_{r-1})$ has a limit and the limit is what we mean by the area A under the curve between a and b. The limit is called the integral:

DEFINITION: *If $f(x)$ is bounded on $[a, b]$, where $a < b$, and if the limit process $F(P)$ of the sum $\sum_{r=1}^{n} f(x_r')(x_r - x_{r-1})$ converges over stages P as the partition P is refined, then the* **integral** *of $f(x)$ from a to b exists:*

$$\int_a^b f(x)\,dx = \operatorname*{Lim}_{P} \sum_{r=1}^{n} f(x_r')(x_r - x_{r-1}).$$

If $f(x) \geqslant 0$ *on* $[a, b]$, *then* $\int_a^b f(x)\,dx$ *is the* **area** *between the curve* $y = f(x)$ *and the axis Ox and between* $x = a$ *and* $x = b$ (Fig. 10.5*a*).

In amplification of this definition, which is in purely algebraic and limit terms, it is to be stressed that $\int_a^b f(x)\,dx$ depends on the form of the function and on the values assigned to a and b, but *not* on the variable x itself. The variable is 'integrated out' over the interval $[a, b]$. Hence

$$\int_a^b f(x)\,dx = \int_a^b f(u)\,du = \int_a^b f(t)\,dt = \dots$$

being always the same area or integral whatever label is given to the variable. The value of the integral only changes if a different function g replaces f or if a different interval $[c, d]$ replaces $[a, b]$.

The notation $\int_a^b f(x)\,dx$ is perhaps not the best which could be devised. Indeed, when the integral is viewed as an operator applied to the function $f(x)$, a quite different notation is later introduced. But $\int_a^b f(x)\,dx$ is the notation in common use and its origin is as follows. The sum $\sum_{r=1}^{n} f(x_r')(x_r - x_{r-1})$ means select x_r' in the segment $(x_r - x_{r-1})$, form the product shown and add for all segments from a to b. It is sometimes written $\overset{b}{\underset{a}{S}} f(x)\Delta x$, where x is selected in a segment of length Δx around x, the product formed and the sum taken for all segments Δx adding to $b - a$. In the limit $\overset{b}{\underset{a}{S}} f(x)\Delta x \rightarrow \int_a^b f(x)\,dx$. This is only dangerous if the meaningless part '$f(x)\,dx$' of the notation is separated off to be read as the value of the function $f(x)$ times a 'small increment' or 'differential' dx.

The evaluation of the integral of particular functions, and the establishment of properties of integrals, from the definition is an extremely tricky and laborious business. Even if we are prepared to assume that the integral exists (which is in fact the case for any continuous function, as stated in 10.6 below), we are still left with the laborious method of finding the value of the integral as a sequential limit, for a partitioning of $[a, b]$ into n equal segments (10.4). This is not a practical proposition. We abandon it right away and look for something quicker. Fortunately, we find something better in 10.7 below.

Meanwhile we can write down a few properties and particular integrals which can be established without too much trouble from the basic definition. First, proceeding on the lines of 10.4 (see 10.9 Ex. 17), we get:

$$\int_a^b k\,dx = k(b-a) \text{ (k constant)}; \int_a^b x\,dx = \tfrac{1}{2}(b^2-a^2) \quad \dots\dots\dots(3)$$

Next, the following simple but useful properties follow for functions $f(x)$ and $g(x)$ which have integrals:

$$\int_a^b \{k f(x)\}\,dx = k\int_a^b f(x)\,dx \quad \text{(k constant)}$$

$$\int_a^b \{f(x)+g(x)\}\,dx = \int_a^b f(x)\,dx + \int_a^b g(x)\,dx.$$

But nothing, as yet, can be offered for the integral of a product (or quotient) of two functions. To illustrate, a sketch of the proof of the sum property is given: use the same partition P of $[a, b]$ for both $f(x)$ and $g(x)$, writing $\sum_{r=1}^{n} f(x_r')(x_r - x_{r-1})$ as contained in $[L_1(P), G_1(P)]$ and $\sum_{r=1}^{n} g(x_r')(x_r - x_{r-1})$ as contained in $[L_2(P), G_2(P)]$. Then $\sum_{r=1}^{n} \{f(x_r') + g(_r')\}(x_r - x_{r-1})$ is contained in the interval $[L(P), G(P)]$ where $L(P) = L_1(P) + L_2(P)$ and $G(P) = G_1(P) + G_2(P)$. Taking the limit over stages P, the result for $\int_a^b f(x) + g(x)\,dx$ follows.

Finally, as a property of a different kind:

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx \quad (a<c<b) \quad \dots\dots\dots\dots(4)$$

To prove this, it is only necessary to concentrate on all those partitions P which include the given c as one of the dividing points; all sums then split into two, one part for the interval $[a, c]$ and the other for $[c, b]$, giving the result shown.

It is not convenient in practice to confine integrals to intervals $[a, b]$ which are such that $a<b$. We often need to write $\int_a^b f(x)\,dx$ even if $a=b$ or $a>b$. This is only a matter of adopting appropriate conventions:

$$\int_a^a f(x)\,dx=0 \quad \text{and} \quad \int_a^b f(x)\,dx=-\int_b^a f(x)\,dx.$$

The first says that the area on an interval of zero length is zero. The second convention is that an area from right to left, on the interval $[a, b]$ which runs from right to left ($a>b$), is equal in numerical value but opposite in sign to the corresponding area on the interval $[b, a]$ from left to right.

Some matters concerned with the use of integrals as areas can be cleared up. We need to check that, as particular cases, the areas of rectangles and triangles given by integrals agree with the elementary notions of these areas. Consider the rectangle $OLQN$, with sides a and k, shown in Fig. 10.5c. This is the area under the line $y=k$, from $x=0$ to $x=a$: $\int_0^a k\,dx=ka$ by (3), i.e. area is base × height as required.

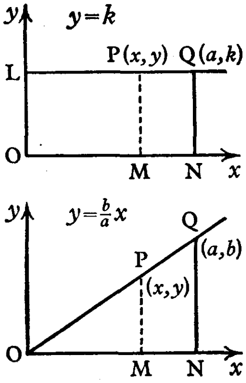


FIG. 10.5c

Consider, further, the triangle OQN of Fig. 10.5c, with base a and height b. This is the area under the line $y=\frac{b}{a}x$ from $x=0$ to $x=a$. For, if $P(x, y)$ is a point on the line, then:

$$\frac{y}{x}=\frac{MP}{OM}=\frac{NQ}{ON}=\frac{b}{a} \quad \text{i.e. } y=\frac{b}{a}x.$$

Hence, area $=\int_0^a \frac{b}{a}x\,dx=\frac{b}{a}\int_0^a x\,dx=\frac{b}{a}(\tfrac{1}{2}a^2)=\tfrac{1}{2}ab$ by (3).

Again, the area is as required: $\frac{1}{2}$ base × height.

The integral and area $\int_a^b f(x)\,dx$ is the limit of $\sum_{r=1}^{n} f(x_r')(x_r-x_{r-1})$, a purely algebraic concept. An important matter of signs has to be considered. If $f(x) \geqslant 0$ everywhere over the interval $[a, b]$, then the sum, and its limit the integral, is essentially positive. $\int_a^b f(x)\,dx$ is the area A of Fig. 10.5a. But, if $f(x) < 0$ anywhere over the interval, then there are negative as well as positive terms in the sum $\sum_{r=1}^{n} f(x_r')(x_r - x_{r-1})$ and the algebraic sum (and hence the integral as limit) is the net balance of positive and negative terms. The position can be represented as in Fig. 10.5d. Suppose $f(x) < 0$ everywhere over $[a, b]$. Then the sum consists entirely of negative terms and the area (of numerical value B) is entirely below Ox: $B = -\int_a^b f(x)\,dx$. However, suppose the curve $y = f(x)$ crosses Ox at $x = c$, such that $f(x) > 0$ for $a \leqslant x < c, f(x) = 0$ at $x = c, f(x) < 0$ for $c < x \leqslant b$. Then the sum consists partly of positive and partly of negative terms and the integral is the net balance between the areas C above Ox and D below Ox shown in the diagram. To get the numerical values (C and D) of the areas, we split the interval into two parts at c and we use the result (4) above:

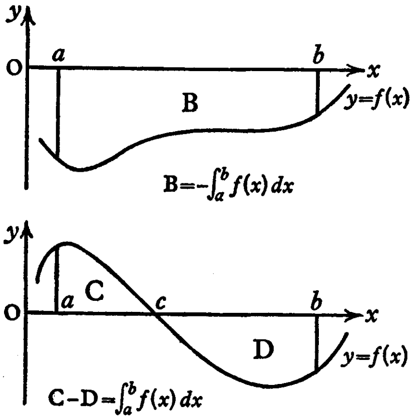


FIG. 10.5d

$$C = \int_a^c f(x)\,dx \quad \text{and} \quad D = -\int_c^b f(x)\,dx$$

i.e. the total numerical area is:

$$\int_a^c f(x)\,dx - \int_c^b f(x)\,dx = C + D$$

whereas the total integral is:

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx = C - D.$$

In using integrals, particularly with reference to the evaluation of areas, we must make allowance for the fact that they are algebraic, and not numerical, sums.

10.6. The fundamental theorem of the calculus. The theorem now to be stated and applied is one of the most extraordinary in all mathematics. It is simple enough to state and it says two things. One is that the integral of a continuous function always exists, a result which is, to say the least, extremely convenient and useful. The other is that integration is the inverse operation to derivation. It is this which gives the theorem its great power; it goes right down to the fundamentals of the calculus.

Some preliminary considerations will help us to appreciate this powerful theorem. First we need to get an integral into the form of a function of x. So far, the integral and area are written $\int_a^b f(x)\,dx$, depending on a and b. Write:

$$F(x)=\int_a^x f(u)\,du \quad \text{..........................}(1)$$

where the variable of integration (which is a matter only of labelling) is written as u to prevent any ambiguity with the variable x, which now stands for the upper end of the interval $[a, x]$ over which the integral is obtained. Then $F(x)$ is an integral as a function of x, as required. We must remember, however, that the value of a, the lower or fixed end of the interval $[a, x]$ used, also enters into $F(x)$.

Suppose that we have two functions, $F(x)$ and $f(x)$, related in such a way that $F'(x)=f(x)$. In other words, $f(x)$ is the derivative of $F(x)$ and (if it exists at all) $f(x)$ is uniquely obtained from $F(x)$. The curve $y=F(x)$ has a tangent at x with slope equal to the height of the curve $y=f(x)$. Or, what comes to the same thing, $F(x)$ is an anti-derivative of $f(x)$ and, as such, is subject to the addition of an arbitrary constant (10.3). The suggestion of the example (i) of 10.1 is that the area under the curve $y=f(x)$ on the interval $[a, x]$ is equal to the height of the curve $y=F(x)$ at x. If it is true, then $\int_a^x f(u)\,du=F(x)$ where $F'(x)=f(x)$, i.e. integration is the inverse of derivation and the integral (1) is the same as an anti-derivative $F(x)$ of $f(x)$. To establish this relation is the job of the Fundamental Theorem of the Calculus.

There are, then, some further properties which we know or expect to hold. In derivation, we pass from $F(x)$ to $f(x)$ where $F'(x)=f(x)$.

We know we cannot count on getting $f(x)$ from $F(x)$, even if $F(x)$ is continuous. We must keep an eye open for the odd case where a continuous $F(x)$ fails to have a derivative, i.e. where the curve $y=F(x)$ has one or more sharp points. In integration, we are proceeding the other way, from $f(x)$ to its integral or anti-derivative $F(x)$. Here we expect that, if $f(x)$ is continuous, then $F(x)$ exists. Our expectation is based on intuition; it seems right that an area exists under a curve which has no discontinuities. Again, it is the job of the Fundamental Theorem to establish that our intuitive conclusion is correct.

Further, any additive constant disappears in derivation. If $F'(x)=f(x)$, then $G(x)=F(x)+A$ has the same derivative: $G'(x)=f(x)$. The inverse process of finding an anti-derivative $F(x)$ of $f(x)$ is always subject to the re-introduction of an arbitrary constant. We can only say that $F(x)$ is *an* anti-derivative; the general form of the anti-derivative includes an additive constant. When the integral (1) is associated with the anti-derivative of $f(x)$, the question of how to incorporate the additive constant remains. The Fundamental Theorem shows how this is related to the fixing of the lower end a of the interval $[a, x]$ used in (1).

FUNDAMENTAL THEOREM OF THE CALCULUS: *If $f(x)$ is defined and continuous on the interval $[a, b]$, then at each x of $[a, b]$: $F(x)=\int_a^x f(u)\,du$ exists, is continuous and has derivative $F'(x)=f(x)$.*

COROLLARY: *If $f(x)$ is continuous with anti-derivative $F(x)$ at each x of $[a, b]$, then: $\int_a^b f(x)\,dx=F(b)-F(a)$.*

The proof of the theorem is difficult. It is set out in 15.5 but omitted here. The corollary follows easily: by the theorem, $\int_a^x f(u)\,du$ and $F(x)$ have the same derivative $f(x)$ and so differ only by a constant. Hence:

$$\int_a^x f(u)\,du=F(x)+A.$$

Put $x=a$: $\qquad 0=F(a)+A \quad$ i.e. $A=-F(a)$.

Put $x=b$: $\qquad \int_a^b f(u)\,du=F(b)+A=F(b)-F(a)$.

Hence: $$\int_a^b f(x)\,dx = F(b) - F(a). \qquad \text{Q.E.D.}$$

A convenient exposition and a suitable notation for integration as inverse derivation can now be set out. Given only that $f(x)$ is defined and continuous on $[a, b]$, the Fundamental Theorem states that an anti-derivative of $f(x)$ exists. Let $F(x)$ be any form of it, i.e. any function found to be such that $F'(x) = f(x)$. Then the general form is $G(x) = F(x) + A$, where A is some constant. The integral of $f(x)$ can be written in three ways in terms of $F(x)$ where $F'(x) = f(x)$:

Firstly:
$$\left.\begin{aligned} &\int_a^b f(x)\,dx = F(b) - F(a) \\ \text{or}\quad &\int_a^b F'(x)\,dx = F(b) - F(a) \end{aligned}\right\}\text{.........................(2)}$$

The arbitrary element does not appear in this form; it goes when the interval $[a, b]$, over which the integral (2) is written, is fixed. For, if $G(x) = F(x) + A$, then:

$$G(b) - G(a) = \{F(b) + A\} - \{F(a) + A\} = F(b) - F(a).$$

The form (2) holds whatever anti-derivative F is taken.

Secondly:
$$\left.\begin{aligned} &\int_a^x f(u)\,du = F(x) - F(a) \\ \text{or}\quad &\int_a^x F'(u)\,du = F(x) - F(a) \end{aligned}\right\}\text{.........................(3)}$$

Allowance is again made for the arbitrary element, in the specification of the fixed end a of the interval $[a, x]$, over which the integral (3) is written. The arbitrary constant appears as $A = -F(a)$, i.e. it is a which is arbitrary in (3).

Thirdly:
$$\left.\begin{aligned} &\int f(x)\,dx = F(x) + A \\ \text{or}\quad &\int F'(x)\,dx = F(x) + A \end{aligned}\right\}\ (A \text{ arbitrary constant}) \text{ ....(4)}$$

can be used as a convenient notation. It implies that the integral is written over some interval $[a, x]$ where x is the variable of the function $F(x)$ and where a is arbitrary and absorbed into the constant $A = -F(a)$.

Of these notations, (2) is called the *definite integral* and (4) is the *indefinite integral*. The link between them is (3) which shows how the arbitrary element is switched from A to a by means of $A = -F(a)$.

In the indefinite integral, there is no function of x. The value $F(b)-F(a)$ shows the dependence of the integral on the interval $[a, b]$ over which it is taken. In the indefinite integral, the emphasis is on the function $F(x)$. The variable x is the upper end of the interval $[a, x]$ over which the integral is taken. There is an additive constant A to allow for an arbitrary lower end a of the interval.

10.7. Integration in practice. The position is that we have abandoned, as too laborious in practice, the evaluation of the definite integral $\int_a^b f(x)\,dx$ as a limit. Instead we concentrate on the indefinite integral $\int f(x)\,dx=F(x)+A$, where $F(x)$ is some anti-derivative of $f(x)$ and where A is an arbitrary constant. We seek a function $F(x)$ such that $F'(x)=f(x)$, knowing at least that it must exist if $f(x)$ is continuous. Once $F(x)$ is found, there is no difficulty whatever in evaluating a definite integral of $f(x)$:

$$\text{If } \int f(x)\,dx=F(x)+A, \text{ then } \int_a^b f(x)\,dx=\Big[F(x)\Big]_a^b=F(b)-F(a)$$

and the arbitrary constant disappears.

Note that we seek *an* anti-derivative not *the* anti-derivative. There are various anti-derivatives and one differs from another by a constant. This is not quite as trivial as it might seem. For example:

$$D\{x(x+2)\}=D(x^2+2x)=D(x^2)+2D(x)=2x+2=2(x+1)$$

i.e.

$$\int 2(x+1)\,dx=x(x+2)+A$$

and an anti-derivative of $2(x+1)$ is $x(x+2)$. We could equally well have found $(x+1)^2$ or $(x-1)(x+3)$. Though these look different, they do in fact differ from $x(x+2)$ only by a constant. We should never be surprised to find two apparently different anti-derivatives of a given function on offer; they can be perfectly respectable in that they differ by a constant.

Viewed in this way, integration is just reverse derivation. However, we shall see that the rules of the game are a little more difficult for integration. The set of operational rules for derivatives (10.3) do not provide so tidy a set of rules for integrals. Essentially, integration is a hit-or-miss affair: just find somehow an anti-derivative $F(x)$ such that $F'(x)=f(x)$, add an arbitrary constant for the indefinite integral

$\int f(x)\,dx$, write $F(b) - F(a)$ for the definite integral $\int_a^b f(x)\,dx$, and we are done. As practical guides, without expecting too much, we can see what we can do in inverting the rules and standard forms for derivatives.

In the following development, the indefinite integrals shown are assumed to exist. They are written without the additive constant for convenience; but it must always be borne in mind. Standard forms for derivatives can always be inverted to give standard forms for integrals. The one we need immediately, for algebraic functions, is the inversion of $D\left(\frac{x^{r+1}}{r+1}\right) = \frac{(r+1)x^r}{r+1} = x^r$:

$$\int x^r\,dx = \frac{x^{r+1}}{r+1} \quad (r \text{ rational, } r \neq -1) \ldots\ldots\ldots\ldots(1)$$

Of the rules, we can easily handle those for a sum, a difference, a multiplicative constant and an additive constant. If u and v are two continuous functions:

$\int (u \pm v)\,dx = \int u\,dx \pm \int v\,dx$ matching $D(u \pm v) = Du \pm Dv$;

$\int ku\,dx = k\int u\,dx$ matching $D(ku) = kDu$

and $\int (u + A)\,dx = \int u\,dx + Ax$

since $\int (u+A)\,dx = \int u\,dx + A\int 1\,dx = \int u\,dx + Ax$ by the sum and multiplicative constant rules and by uses of (1) with $r = 0$.

The rule for the derivative of a product is itself not a very simple one: $D(uv) = uDv + vDu$. It is not possible to invert it into a corresponding form for integrals. The nearest we can get is a formula, much used in practical integration, called 'integration by parts'. Given two functions u and v with derivatives u' and v', start from the property (4) of 10.6:

$$uv = \int D(uv)\,dx = \int (uv' + vu')\,dx = \int uv'\,dx + \int vu'\,dx.$$

Hence, the formula for *integration by parts:*

$$\int uv'\,dx = uv - \int vu'\,dx \quad \ldots\ldots\ldots\ldots(2)$$

An alternative expression is obtained by writing $u = f(x)$, $u' = f'(x)$, $v' = g(x)$ and $v = \int g(x)\,dx$. So:

$$\int f(x)g(x)\,dx = f(x)\int g(x)\,dx - \int \{f'(x)\int g(x)\,dx\}\,dx \ldots\ldots\ldots(3)$$

The formula (2) or (3) does not serve directly to integrate a product of two functions. It simply passes the buck. The hope, quite often realised in practice with ingenious selection of u and v, is that the integrals on the right-hand side of (2) or (3) are easier to evaluate. An example illustrates:

From (1), we have $\int \sqrt{x}\,dx = \int x^{\frac{1}{2}}\,dx = \frac{x^{\frac{1}{2}+1}}{\frac{1}{2}+1} = \frac{2}{3}x\sqrt{x}$

and $\int \frac{1}{\sqrt{x}}\,dx = \int x^{-\frac{1}{2}}\,dx = \frac{x^{-\frac{1}{2}+1}}{-\frac{1}{2}+1} = 2\sqrt{x}.$

Write $f(x) = 1+3x$ and $g(x) = \frac{1}{2\sqrt{x}}$ in (3):

$$\begin{aligned}\int \frac{1+3x}{2\sqrt{x}}\,dx &= (1+3x)\int \frac{1}{2\sqrt{x}}\,dx - \int\left\{D(1+3x)\times\int\frac{1}{2\sqrt{x}}\,dx\right\}dx\\ &= (1+3x)\tfrac{1}{2}(2\sqrt{x}) - \int\{3\times\tfrac{1}{2}(2\sqrt{x})\}\,dx\\ &= (1+3x)\sqrt{x} - 3\int\sqrt{x}\,dx\\ &= (1+3x)\sqrt{x} - 3\tfrac{2}{3}x\sqrt{x}\\ &= (1+x)\sqrt{x}.\end{aligned}$$

As a check: $D\{(1+x)\sqrt{x}\} = \frac{1+3x}{2\sqrt{x}}$ as shown in 10.3.

The other rule of derivation, of frequent use in practice, is that for a composite function. If y is a function of u and u a function of x:

$$D_xy = D_uyD_xu.$$

Something can be done with this to give a practical rule for integration known as 'integration by substitution'. Let $y=f(u)$, where $u=g(x)$ with derivative $u'=g'(x)$. Write:

$$\int f(u)\,du = F(u) = F\{g(x)\}$$

treating it either as a function of u or as a composite function of x. As a function of u, $F'(u)=f(u)$. As a function of x, writing $D=D_x$ for derivatives with respect to x, we have:

$$\begin{aligned}DF\{g(x)\} &= D_uF(u)Du \quad \text{by the composite function rule}\\ &= f(u)u' \quad \text{where } u=g(x),\ u'=g'(x)\end{aligned}$$

i.e. $$F\{g(x)\} = \int f(u)u'\,dx.$$

Hence the formula for *integration by substitution:*

$$\int f(u)\,du = \int f(u)u'\,dx \quad\text{..........................(4)}$$

where u is a function of x with derivative u'. More fully:

$$\int f(u)\,du = \int f\{g(x)\}g'(x)\,dx \quad \ldots\ldots\ldots\ldots\ldots\ldots(5)$$

on making the substitution $u=g(x)$. The use of (4) or (5) is that, to evaluate the integral of $f(u)$, we switch variables from u to x by substituting $u=g(x)$. Again, this is a successful passing of the buck only if the integral obtained on the right-hand side of (4) or (5) is easily evaluated. This hope is often realised by appropriate choice of $g(x)$. The same example illustrates:

$$\int \frac{1+3u}{2\sqrt{u}}\,du = \int \frac{1+3x^2}{2x}\,2x\,dx \quad \text{on substituting } u=x^2,\ u'=2x$$

$$= \int (1+3x^2)\,dx = \int 1\,dx + 3\int x^2\,dx = x + x^3 \quad \text{by (1).}$$

Substituting back $u=x^2$: $\displaystyle\int \frac{1+3u}{2\sqrt{u}}\,du = (1+u)\sqrt{u}$ as before.

In integration, there may be more than one way of skinning a cat.

The rules for integration can be brought together. Here u and v are continuous functions with integrals. To ensure that the integrals exist in integration by parts, the derivatives u' and v' are also taken as continuous. In integration by substitution, $f(u)$ and $u'=g'(x)$ are both taken as continuous.

Rule	*Indefinite Integral**
Additive constant	$\int (u+A)\,dx = \int u\,dx + Ax$ (A constant)
Sum	$\int (u+v)\,dx = \int u\,dx + \int v\,dx$
Difference	$\int (u-v)\,dx = \int u\,dx - \int v\,dx$
Multiplicative constant	$\int ku\,dx = k\int u\,dx$ (k constant)
Integration by parts	$\int uv'\,dx = uv - \int vu'\,dx$
Integration by substitution	$\int f(u)\,du = \int f(u)u'\,dx$ where $u=g(x)$

Finally, to stress the connection between indefinite and definite integrals, consider the particular example already used:

$$\int \frac{1+3x}{2\sqrt{x}}\,dx = (1+x)\sqrt{x}.$$

In writing an indefinite integral such as this, we indicate that we have a function of a variable x by using x after the $\int$ sign. If we change

* An arbitrary constant is to be added in each case.

the variable after the $\int$ sign, we keep the same function but change the independent variable to which it relates. So:

$$\int \frac{1+3x}{2\surd x}\,dx = (1+x)\surd x \quad \text{and} \quad \int \frac{1+3u}{2\surd u}\,du = (1+u)\surd u$$

the first a function of x, the second a function (the same one) of u. For a definite integral, the 'variable' used after the $\int$ sign does not imply a function of this 'variable'. It is a 'dummy variable', to be labelled in any way convenient. For, the definite integral involves only the ends of the interval $[a, b]$ over which it is taken; it is not a function of a variable x. So:

$$\int_a^b f(x)\,dx = \int_a^b f(u)\,du = \ldots = F(b) - F(a)$$

where F is an anti-derivative of f. In the particular case:

$$\int_1^4 \frac{1+3x}{2\surd x}\,dx = \Big[(1+x)\surd x\Big]_1^4 = \Big[(1+x)\surd x\Big]_{x=4} - \Big[(1+x)\surd x\Big]_{x=1} = 10-2$$

i.e. $$\int_1^4 \frac{1+3x}{2\surd x}\,dx = 8$$

and the same value is obtained if we start with $\displaystyle\int_1^4 \frac{1+3u}{2\surd u}\,du$.

10.8. Derivatives and integrals as operators. When introducing the derivative notation (10.2), we remarked that 'D' in $DF(x)$ could be taken as an operator and read 'get the derivative of'. We have a transformation, changing a given function into another function, i.e. the derived function. This follows the usage of 6.4. Some questions immediately suggest themselves. Can integration also be written in operator form? If so, is the operator the inverse of D? Can a whole group of such operators be written?

In pursuing these matters, we ignore for the moment the arbitrary constants which arise in integration. For some continuous $f(x)$, write $\int f(x)\,dx = F(x)$, meaning $DF(x) = f(x)$. Let the transformation from $f(x)$ to an integral or anti-derivative $F(x)$ be written in operator form: $Ef(x) = F(x)$. Here 'E' is read 'integrate' or 'get an anti-derivative of'; $Ef(x) = F(x)$ means that getting an anti-derivative of $f(x)$ produces $F(x)$. Then E and D are easily related:

If $Ef(x) = F(x)$, so that $DF(x) = f(x)$, then $D\{Ef(x)\} = DF(x) = f(x)$.

In the usual notation for transformations, the product DE is written for successive applications of the operators, E first and D second. So: $DEf(x) = f(x)$. Further, $E\{DF(x)\} = Ef(x) = F(x)$, i.e. $EDF(x) = F(x)$ where ED is the succession of D first and E second. Introduce I for the identity operator, leaving a function unchanged: $If(x) = f(x)$. We now have three operators to apply one after the other and they are related: $DE = ED = I$. Hence D and E are commutative in multiplication, interpreted as successive application, and one is the inverse of the other. We can write E as D^{-1}, the inverse of D:

$$DD^{-1} = D^{-1}D = I.$$

We now have a new notation for integration, in many ways more convenient than the old: $D^{-1} = \int \dots\, dx$. So: $\int f(x)\, dx = D^{-1}f(x)$.

The idea of successive application of the operators D and D^{-1} can be pursued further. Suppose $F(x)$ has derivative $F'(x)$ and suppose that this derived function has a derivative in its turn. This is the *second derivative* of $F(x)$, written $F''(x)$. If $F'(x)$ is interpreted as the 'velocity' of $F(x)$, then $F''(x)$ is the 'acceleration' (see 10.9 Ex. 7). In operator form:

$$F''(x) = DF'(x) = DDF(x) = D^2F(x).$$

If each derived function always has a derivative, the process continues:

DEFINITION: *If $F(x)$ has derivative $F'(x)$, if $F'(x)$ has derivative $F''(x)$, ... for n stages (n a positive integer), then there is an* **nth derivative:**

$$D^nF(x) = F^{(n)}(x).$$

Suppose $f(x)$ is continuous so that $\int f(x)\, dx$ exists and is continuous. Then $\int f(x)\, dx$ has an integral, the *second integral* of $f(x)$, written $\int\int f(x)\, dx\, dx$:

$$\int\int f(x)\, dx\, dx = D^{-1}\int f(x)\, dx = D^{-1}D^{-1}f(x) = D^{-2}f(x).$$

Write $F(x) = \int f(x)\, dx$ and $G(x) = \int F(x)\, dx = \int\int f(x)\, dx\, dx$. Then $DF(x) = f(x)$ and $DG(x) = F(x)$. Hence $D^2G(x) = DF(x) = f(x)$. But $G(x) = D^{-2}f(x)$. Hence $D^{-2}f(x) = G(x)$ implies $D^2G(x) = f(x)$, i.e.

$$D^2D^{-2}f(x) = f(x) \quad \text{and} \quad D^{-2}D^2G(x) = G(x).$$

So D^{-2} and D^2 are inverse: $D^2D^{-2} = D^{-2}D^2 = I$.

Since $f(x)$ is continuous, the process of successive integration continues:

DEFINITION: *If $f(x)$ is continuous, the* **nth integral**

$$D^{-n}f(x)=\int\int \dots \int f(x)\,dx\,dx \dots dx$$

exists for any positive integer n, and D^{-n} is the inverse of D^n.

As obvious conventions, write $D^0=I$ and $D^1=D$. A general operator D^n is obtained for any integer n, positive, zero or negative:

$$\dots D^{-2}, D^{-1}, I, D, D^2, \dots$$

as a set of operators or transformations. If n is positive, then $D^n=D\times D\times\dots\times D$ (n times) for the nth derivative. If n is negative, write it $-m$ so that $D^{-m}=D^{-1}\times D^{-1}\times\dots\times D^{-1}$ (m times) for the mth integral. Further, the operators can be mixed, and are commutative, in successive applications (10.9 Ex. 29). In general:

$$D^mD^n=D^{m+n} \quad (m \text{ and } n \text{ integers}).$$

The set of operators is a group under multiplication, with an identity $I=D^0$ and with every member of the set having an inverse: $D^nD^{-n}=I$. There is a one-one correspondence between the set D^n and the integers n. The group D^n under multiplication is isomorphic with the group of integers under addition.

There are now qualifications to be noted. Run through the sequence of higher and higher derivatives of a given function $F(x)$:

$$F(x), DF(x), D^2F(x), \dots \quad \text{or} \quad F(x), F'(x), F''(x), \dots .$$

This can only be done if the function at each stage does have a derivative. The sequence, in fact, may be halted at any stage for lack of a derivative. To write the sequence up to a stage n, we must assume (or ensure) that the first n derivatives of $F(x)$ exist. On the other hand, given a continuous function $f(x)$, we can always write the sequence:

$$f(x), D^{-1}f(x), D^{-2}f(x), \dots \quad \text{or} \quad f(x), \int f(x)\,dx, \int\int f(x)\,dx\,dx, \dots .$$

The difficulty here is a different one. At each stage, an arbitrary constant is introduced so that, by the nth stage, there are n of them:

$$\begin{aligned}
D^{-1}f(x)&=\int f(x)\,dx+A_1\\
D^{-2}f(x)&=\int\int f(x)\,dx\,dx+A_1x+A_2\\
D^{-3}f(x)&=\int\int\int f(x)\,dx\,dx\,dx+A_1\frac{x^2}{2}+A_2x+A_3\\
&\dots\dots\dots\dots\dots\dots
\end{aligned}$$

As a simple example, take the function x^2. Successive derivatives do exist:

$$x^2;\ Dx^2 = 2x;\ D^2x^2 = 2;\ D^3x^2 = D^4x^2 = \ldots = 0.$$

Successive integrals can always be written but they involve a mounting set of additive constants:

$$x^2;\ D^{-1}x^2 = \frac{x^3}{3} + A_1;\ D^{-2}x^2 = \frac{x^4}{3\,.\,4} + A_1x + A_2;$$

$$D^{-3}x^2 = \frac{x^5}{3\,.\,4\,.\,5} + A_1\,\frac{x^2}{2} + A_2x + A_3;\ \ldots$$

In using the operator form, we can omit the arbitrary constants, e.g. $D^{-3}x^2 = \dfrac{x^5}{3\,.\,4\,.\,5}$, as long as we remember the qualification that they need to be inserted to get from *an* integral to the *general* integral.

10.9. Exercises.

1. A body travels $x = 100t - \frac{1}{3}t^3$ feet in t minutes. Five minutes have elapsed; find the average speed over 1 minute more, over 0·1 minute, over 0·01 minute. Write the average speed over h minutes and deduce the velocity after 5 minutes. Generally, show $v = 100 - t^2$ and deduce that the formula for x applies to a body starting off at 100 feet per minute, with decreasing velocity, coming to rest after 10 minutes.

2. As in 9.9 Ex. 5, show that $y = \dfrac{(x+a)^2 - a^2}{x} = x + 2a$ $(x \quad 0)$ and $y \to 2a$ as $x \to 0$. Interpret as the derivative of $y = x^2$ at $x = a$.

3. Show that $f(x) = 1 - \surd\{(1-x)^2\}$ is continuous at $x = 1$. Show that

$$\phi(h) = \frac{f(1+h) - f(1)}{h}$$

is $+1$ (h negative) or -1 (h positive). Deduce that the limit process for $\phi(h)$ as $h \to 0$ is $F(N) = [-1, 1]$ for all neighbourhoods N, i.e. that $\phi(h)$ has no limit. Represent $f(x)$ graphically as a 'curve', with a sharp point where the function is continuous but without derivative. Show that $f'(x) = 1$ for $x < 1$, $f'(x)$ not defined at $x = 1$ and $f'(x) = -1$ for $x > 1$.

4. *Tangent and normal.* From (3) of 8.6, show that the tangent at P (x_1, y_1) to the curve $y = f(x)$ has equation $y - y_1 = f'(x_1)(x - x_1)$ if the tangent exists, and that the line perpendicular to the tangent at P (called the *normal* at P) is $(x - x_1) + f'(x_1)(y - y_1) = 0$. Write the equation of the tangent and normal to the parabola $y = x^2$ at the point where $x = x_1$.

5. Given $D(x) = 1$ and $D(x^n) = nx^{n-1}$, use the product rule to establish that $D(x^{n+1}) = (n+1)x^n$. Hence prove $D(x^n) = nx^{n-1}$ by mathematical induction (n a positive integer). From $D\left(\dfrac{1}{x}\right) = -\dfrac{1}{x^2}$, show $D\left(\dfrac{1}{x^n}\right) = -\dfrac{n}{x^{n+1}}$ by the composite function rule.

6. It is given that $D(x^n) = nx^{n-1}$, n a positive integer. Take p and q as positive integers. Show that $D(\sqrt[q]{x}) = 1/q \sqrt[q]{(x^{q-1})}$ by the inverse function rule. Use the composite function rule to deduce $D(\sqrt[q]{x^p})$.

7. *Second derivative as an acceleration.* Distance travelled in time t is given as $x = f(t)$ so that velocity $v = f'(t)$. Show that $f''(t)$ is the limit of the average rate of change of v at time t and interpret as acceleration. Show that the acceleration is constant for the motion of (i) of 10.1 and that it is negative (deceleration) in Ex. 1.

8. Obtain $\dfrac{d}{dx}\left(\dfrac{1}{x}\right)$ from first principles by writing $\dfrac{f(x+h) - f(x)}{h} = -\dfrac{1}{x(x+h)}$ for $f(x) = \dfrac{1}{x}$.

9. Establish from the definition the quotient rule for derivatives.

10. Show that the derivatives of $1 + x + x^2$; $\dfrac{1+x^2}{1-x^2}$; $\dfrac{x + \sqrt{x+1}}{\sqrt{x-1}}$; and $\dfrac{1}{\sqrt{1-x^2}}$ are respectively: $1 + 2x$; $\dfrac{4x}{(1-x^2)^2}$; $\dfrac{(x-2)\sqrt{x+1} - 2}{2(x-1)\sqrt{x^2-1}}$; and $\dfrac{x}{(1-x^2)\sqrt{1-x^2}}$. Show also that $\sqrt{x} + \dfrac{1}{\sqrt{x}}$ and $\dfrac{(1+\sqrt{x})^2}{\sqrt{x}}$ are both anti-derivatives of $\frac{1}{2}\dfrac{x-1}{x\sqrt{x}}$. Check by showing that they differ by a constant.

11. A line has equation $y = mx + c$. Show that $Dy = m$, $D^2y = 0$; and interpret the slope of the line as the slope of a tangent. Conversely, given $D^2y = 0$, write anti-derivatives to obtain $y = mx + c$, i.e. to get $y = mx + c$ as the solution of $D^2y = 0$.

12. If $y = x/(1-x^2)$ show that y, Dy and D^2y are all positive in the open interval $(0 < x < 1)$ and hence that y increases at an increasing rate from $y = 0$ at $x = 0$. What can be said when $x \to 1$? Show also that, if $y = 1 + (2x+3)/(x^2-1)$, then Dy is zero at $x = -2{\cdot}62$ and $x = -0{\cdot}38$ (approximately), positive between these values and negative elsewhere. (See Fig. 3.9.)

13. *A freely-falling body.* A body falls freely from rest; Galileo's law is that $x = \frac{1}{2}gt^2$ (g the *gravitational constant*) is the distance travelled in time t. Show that velocity v increases steadily, with constant acceleration g. Conversely, if the acceleration is given as a constant g, show that $v = u + gt$ and $x = ut + \frac{1}{2}gt^2$, where u is the initial velocity (at $t = 0$, $x = 0$). Further, if the body is thrown *upwards* with velocity u, show that the upward velocity v and the distance x travelled upwards after time t are $v = u - gt$, $x = ut - \frac{1}{2}gt^2$; and deduce that the greatest height obtained is $u^2/2g$.

*14. A ball is thrown into the air with velocity u at an angle $\alpha°$ to the horizontal. Show that horizontally (with no acceleration) the distance travelled is $x = ut \cos \alpha$, and vertically (with downward acceleration g) it is $y = ut \sin \alpha - \frac{1}{2}gt^2$, after time t. Deduce that the ball returns to the ground a distance $x = \dfrac{u^2}{g} \sin 2\alpha$ from the starting point ($\sin 2\alpha = 2 \sin \alpha \cos \alpha$), and that a maximum distance is achieved if the ball is thrown at 45°. Show that the equation of the path of the ball through the air is obtained (by eliminating t) as $y = x \tan \alpha - \frac{g}{2u^2 \cos^2 \alpha} x^2$, which is a parabola.

15. *Marginal revenue.* The demand for the product of a firm (x thousand items per week) at specified prices (£p per item) is given by $x + \frac{1}{2}p = 1$. Write R (in £ thousand) as the total revenue obtained by selling x (thousand items) and show that $\frac{dR}{dx} = 2(1 - 2x)$. Interpret this as *marginal revenue*; what units does it appear in? Put $\frac{dR}{dx} = 0$, achieved at output of 500 items per week, and interpret graphically.

16. *Maximum profits.* In the firm of Ex. 15, the total cost C of output x (thousand items) is $C = 2x^2$ (£ thousand), and marginal cost is $\frac{dC}{dx} = 4x$. Write $\Pi = R - C$ for profits at output x and show that Π is greatest at the output (250 items per week) where marginal cost = marginal revenue.

17. Establish the following areas (definite integrals) as limits of sequences as in 10.4:

$$\int_1^2 1\, dx = 1; \int_1^2 x\, dx = \frac{3}{2}; \int_1^2 x^2\, dx = \frac{7}{3}; \int_1^2 x^3\, dx = \frac{15}{4}.$$

Generalise by using the standard form for $\int x^r\, dx$:

$$\int_a^b 1\, dx = b - a; \int_a^b x\, dx = \tfrac{1}{2}(b^2 - a^2); \int_a^b x^2\, dx = \tfrac{1}{3}(b^3 - a^3); \int_a^b x^3\, dx = \tfrac{1}{4}(b^4 - a^4).$$

18. For $y = 1 - \sqrt{(1 - x)^2}$ defined on $0 \leqslant x \leqslant 2$ (see Ex. 3), obtain

$$\int_0^2 y\, dx = \int_0^1 y\, dx + \int_1^2 y\, dx = 1$$

and check graphically by means of areas of triangles.

19. If $y = x^3 - 3x^2 + 2x$ is represented by a curve, as in Fig. 10.5*d*, interpret $\int_0^2 y\, dx = 0$ in terms of areas under the curve. If C is the area *above* Ox from 0 to 1 and D the area *under* Ox from 1 to 2, show that $C = D = \frac{1}{4}$.

20. From $\frac{x - 1}{x\sqrt{x}} = x^{-1/2} - x^{-3/2}$, show that $\int \frac{x - 1}{x\sqrt{x}}\, dx = 2\left(\sqrt{x} + \frac{1}{\sqrt{x}}\right) + \text{constant}$. Check from Ex. 10.

21. Make the substitution $u = 1 - x^2$ to show that $\int \frac{x\, dx}{(1 - x^2)^2} = \frac{1}{2} \frac{1}{1 - x^2} +$ constant. But $\int \frac{x\, dx}{(1 - x^2)^2} = \frac{1}{4} \frac{1 + x^2}{1 - x^2}$ apart from the additive constant (by Ex. 10); is this consistent?

22. By integration by substitution with $u = x - 1$, show that

$\int \sqrt{x - 1}\, dx = \frac{2}{3}\sqrt{(x - 1)^3}$ and $\int \frac{dx}{\sqrt{x - 1}} = 2\sqrt{x - 1}$ (+ constant in each case).

Then, by integration by parts, show that

$$\tfrac{1}{2}\int \frac{x\,dx}{\sqrt{x-1}} = x\sqrt{x-1} - \tfrac{2}{3}\sqrt{(x-1)^3} + \text{constant}.$$

*23. Show that $\int_0^1 x^{m-1}(1-x)^{n-1}\,dx = \int_0^1 x^{n-1}(1-x)^{m-1}\,dx$ where m and n are positive integers. (Substitute $u = 1 - x$.) Further, show that

$$D\{x^m(1-x)^n\} = mx^{m-1}(1-x)^{n-1} - (m+n)x^m(1-x)^{n-1}$$

and hence that $\int x^m(1-x)^{n-1}\,dx = \dfrac{m}{m+n}\int x^{m-1}(1-x)^{n-1}\,dx - \dfrac{1}{m+n}x^m(1-x)^n$.

Deduce that $\int_0^1 x^m(1-x)^{n-1}\,dx = \dfrac{m}{m+n}\int_0^1 x^{m-1}(1-x)^{n-1}\,dx$

and $\int_0^1 x^{n-1}(1-x)^m\,dx = \dfrac{m}{m+n}\int_0^1 x^{n-1}(1-x)^{m-1}\,dx.$

24. *Convergent infinite integrals.* If $f(x)$ is continuous $x \geqslant a$ and if $\int_a^k f(x)\,dx$ has limit L as $k \to \infty$, define the infinite integral $\int_a^\infty f(x)\,dx = L$ and say that it is convergent. Interpret the convergent infinite integral in terms of areas under the curve $y = f(x)$. Show that $\int_1^\infty \dfrac{dx}{x^2} = 1$ and illustrate graphically.

25. From Ex. 21, show that $\int_h^k \dfrac{x\,dx}{(1-x^2)^2} = \tfrac{1}{2}\left(\dfrac{1}{h^2-1} - \dfrac{1}{k^2-1}\right)$ for $k > h > 1$. Deduce that $\int_2^\infty \dfrac{x\,dx}{(1-x^2)^2}$ is convergent (with value $\tfrac{1}{6}$) but that neither $\int_1^k \dfrac{x\,dx}{(1-x^2)^2}$ nor $\int_1^\infty \dfrac{x\,dx}{(1-x^2)^2}$ can be written. (The integral is not convergent as $h \to 1$.)

*26. *Beta functions.* If m and n are positive integers, write

$$B(m, n) = \int_0^1 x^{m-1}(1-x)^{n-1}\,dx$$

and use the results of Ex. 23 to establish that:

$$B(m, n) = B(n, m) \quad\text{and}\quad B(m+1, n) = \frac{m}{m+n}B(m, n).$$

Deduce that $B(1, 1) = 1,\ B(m, 1) = \dfrac{1}{m},\ B(1, n) = \dfrac{1}{n};$

and that $B(m, n) = \dfrac{(m-1)!\,(n-1)!}{(m+n-1)!} \quad (m > 1, n > 1).$

(The Beta function can be defined similarly for m and n any positive real values.)

* 27. *Transformation of integrals.* The formula for integration by substitution can be regarded as the transformation of one integral into another. Define $B(m, n)$ as the integral of Ex. 26 and transform by $x = \dfrac{1}{1+y}$. Show first that $\int_h^1 x^{m-1}(1-x)^{n-1}\,dx = \int_0^k \dfrac{y^{n-1}}{(1+y)^{m+n}}\,dy$ where $k = \dfrac{1-h}{h} \to \infty$ as $h \to 0$. Deduce that $B(m, n) = \int_0^\infty \dfrac{x^{n-1}}{(1+x)^{m+n}}\,dx$ as a convergent infinite integral.

*28. *Integration by parts for infinite integrals* can be written:

$$\int_a^\infty uv'\,dx = \Big[uv\Big]_a^\infty - \int_a^\infty vu'\,dx \quad \text{where} \quad \Big[uv\Big]_a^\infty = \lim_{x\to\infty}(uv) - \Big[uv\Big]_a.$$

Illustrate its application by showing that:

$$\int_0^\infty \frac{x^{n-1}}{(1+x)^{m+n}}\,dx = \left[\frac{1}{(1+x)^{m+n}}\frac{x^n}{n}\right]_0^\infty + \frac{m+n}{n}\int_0^\infty \frac{x^n}{(1+x)^{m+n+1}}\,dx$$

i.e. that:
$$\int_0^\infty \frac{x^n}{(1+x)^{m+n+1}}\,dx = \frac{n}{m+n}\int_0^\infty \frac{x^{n-1}}{(1+x)^{m+n}}\,dx.$$

Deduce that
$$B(m, n+1) = \frac{n}{m+n}\,B(m, n).$$

Use the symmetry: $B(m, n) = B(n, m)$, to show that this is the same as

$$B(m+1, n) = \frac{m}{m+n}\,B(m, n).$$

29. From the definition of the operator D, show that

$$DDD^{-1}f(x) = DD^{-1}Df(x) = D^{-1}DDf(x) = Df(x).$$

Examine other such combinations, generalise and show that

$$D^m D^n f(x) = D^{m+n} f(x)$$

for any integral m and n.

30. *Leibniz's Theorem.* If the functions u and v have derivatives of any desired order, use $D(uv) = uDv + vDu$ to show that $D^2(uv) = uD^2v + 2DuDv + vD^2u$. Hence prove by mathematical induction, for any positive integer n:

$$D^n(uv) = uD^n v + \binom{n}{1}DuD^{n-1}v + \binom{n}{2}D^2uD^{n-2}v + \ldots + \binom{n}{1}D^{n-1}uDv + vD^n u.$$

CHAPTER 11

EXPANSIONS

11.1. Taylor's series. A function $f(x)$ of a real variable is defined over the interval $[a, b]$. The present development aims at the derivation of an expansion of $f(x)$, approximately as a polynomial in ascending powers of x of degree n, exactly as an infinite series of such powers. The basic theorem, the Mean Value Theorem for derivatives, requires only that $f(x)$ is continuous *over* the whole interval (for $a \leqslant x \leqslant b$) and that $f'(x)$ exists *within* the interval (for $a < x < b$). It is not necessary to assume that $f'(x)$ exists at $x = a$ and at $x = b$ (though it usually does).* Nor is it necessary to assume that $f'(x)$ is continuous (though it usually is). The theorem, reached in two stages, makes use of a theorem on continuity (9.8). Since $f(x)$ is continuous over $[a, b]$, its range is an interval $[c, d]$, every value being achieved for some x of $[a, b]$. In particular, $f(x)$ is bounded and attains its GLB c and its LUB d in the interval.

We begin with what is generally known as *Rolle's Theorem*:

THEOREM: *If $f(x)$ is continuous $a \leqslant x \leqslant b$ and if $f'(x)$ exists $a < x < b$, then $f(a) = f(b)$ implies that α exists $(a < \alpha < b)$ so that $f'(\alpha) = 0$.*

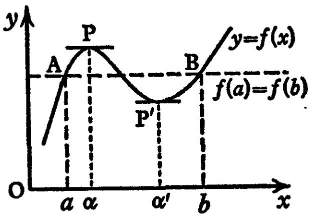


FIG. 11.1*a*

The implication of this result is seen clearly in graphical terms. In a graph of the curve $y = f(x)$ referred to axes Oxy (Fig. 11.1*a*), let A and B be two points at the same height on the curve. Then there is some intermediate point P at which the tangent is horizontal: $f'(\alpha) = 0$. P may not be unique; there can be two (or more) such points, P and P' as illustrated. The theorem states only that there is at least one such point. The proof is as follows:

* If the wider assumption is made that $f'(x)$ exists everywhere over the interval $(a \leqslant x \leqslant b)$, then $f(x)$ is necessarily continuous over the interval.

Write $\phi(x) = f(x) - f(a)$ so that $\phi'(x) = f'(x)$.

Hence: $\phi(a) = \phi(b) = 0$ given $f(a) = f(b)$.

If $\phi(x) = 0$ throughout $[a, b]$, then $\phi'(x) = 0$ everywhere, i.e. $f'(x) = 0$ everywhere and there is nothing more to prove. On the other hand, if $\phi(x) \neq 0$ at some points, there are positive values of $\phi(x)$, or negative values, or both. Suppose $\phi(x) > 0$ somewhere in $[a, b]$. Then the range $[c, d]$ of $\phi(x)$ has $d > 0$, and there is some α $(a < \alpha < b)$ giving the LUB $\phi(\alpha) = d > 0$. It can then be shown that $\phi'(\alpha) = 0$. For, if $\phi'(\alpha) > 0$, $\frac{1}{h}\{\phi(\alpha + h) - \phi(\alpha)\}$ converges to a positive limit as $h \to 0$ and $\frac{1}{h}\{\phi(\alpha + h) - \phi(\alpha)\} > 0$ for some sufficiently small *positive h*. This means that $\phi(\alpha + h) > \phi(\alpha)$ which is ruled out since $\phi(\alpha)$ is the LUB. Equally, if $\phi'(\alpha) < 0$, $\frac{1}{h}\{\phi(\alpha + h) - \phi(\alpha)\} < 0$ for some sufficiently small *negative h*. This means that $\phi(\alpha + h) > \phi(\alpha)$ again, and the case is ruled out. Hence $\phi'(\alpha) = 0$. The same result follows if $\phi(x) < 0$ somewhere in $[a, b]$. Hence, in all cases, $\phi'(\alpha) = 0$, i.e. $f'(\alpha) = 0$ for some α $(a < \alpha < b)$. Q.E.D.

The *Mean Value Theorem for derivatives* then follows:

THEOREM: *If $f(x)$ is continuous $a \leqslant x \leqslant b$ and if $f'(x)$ exists $a < x < b$, then α exists $(a < \alpha < b)$ so that $f'(\alpha) = \dfrac{f(b) - f(a)}{b - a}$.*

The meaning of the result in graphical terms is clear from Fig. 11.1*b*. If A and B are any two points on the curve $y = f(x)$, then there is some point P (perhaps several such) at which the tangent is parallel to the chord AB. The slope of the tangent is $f'(\alpha)$ where $x = \alpha$ corresponds to P. The slope of the chord is:

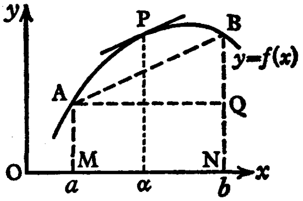


FIG. 11.1*b*

$$\frac{QB}{AQ} = \frac{NB - NQ}{MN} = \frac{NB - MA}{ON - OM} = \frac{f(b) - f(a)}{b - a}.$$

Rolle's Theorem, a particular case, is used to prove the more general case:

Write $$\phi(x) = f(x) - f(a) - (x - a)\,\frac{f(b) - f(a)}{b - a}$$

with $$\phi'(x) = f'(x) - \frac{f(b)-f(a)}{b-a}.$$

But $$\phi(a) = \phi(b) = 0.$$

So, by Rolle's Theorem, there is an α ($a < \alpha < b$) such that $\phi'(\alpha) = 0$,

i.e. such that $$f'(\alpha) - \frac{f(b)-f(a)}{b-a} = 0. \qquad \text{Q.E.D.}$$

Notice that, if $f(x)$ is a specified function, then it may well be possible to locate α as a particular value. For example, if $f(x) = x^2$ with derivative $f'(x) = 2x$, defined over the interval $[1, 2]$, then

$$\frac{f(b)-f(a)}{b-a} = \frac{2^2-1^2}{2-1} = 3 \quad \text{for } a=1,\ b=2$$

and $$f'(x) = 3 \qquad \text{at } x = \tfrac{3}{2}.$$

Hence $$f'(\alpha) = \frac{f(b)-f(a)}{b-a} \quad \text{at } \alpha = \tfrac{3}{2} \text{ in the interval } [a, b] = [1, 2].$$

The importance of the theorem, however, lies in the fact that we know there is a value α whatever function $f(x)$ is taken, subject to the conditions named.

The Mean Value Theorem can be expressed in an equivalent form. Write $b = a+h$ so that $h = b-a > 0$. Then, under the named conditions, we have:

$$f'(\alpha) = \frac{f(a+h)-f(a)}{h} \quad \text{for some } \alpha,\ a < \alpha < a+h.$$

Write $\alpha = a + \theta h$, where θ is a real number such that $0 < \theta < 1$:

$$\frac{f(a+h)-f(a)}{h} = f'(a+\theta h)$$

i.e. $$f(a+h) = f(a) + hf'(a+\theta h).$$

The same result is true if h is negative, say $h = -k$ ($k > 0$). Take the interval in question as $[a-k, a]$ so that the above result modifies to:

$$f(a) = f(a-k) + kf'(a-\theta k)$$

i.e. $$f(a) = f(a+h) - hf'(a+\theta h)$$

i.e. $$f(a+h) = f(a) + hf'(a+\theta h) \quad \text{as before.}$$

An important extension to *Taylor's Theorem* now follows:

THEOREM: *If* $f(x), f'(x), \dots f^{(n)}(x)$ *are continuous* $a \leqslant x \leqslant a+h$ *and if* $f^{(n+1)}(x)$ *exists* $a < x < a+h$, *then*

$$f(a+h)=f(a)+hf'(a)+\frac{h^2}{2!}f''(a)+\ldots+\frac{h^n}{n!}f^{(n)}(a)+R_n$$

where $$R_n=\frac{h^{(n+1)}}{(n+1)!}f^{(n+1)}(a+\theta h)\quad(0<\theta<1).$$

Moreover, as before, the result also holds if h is negative as well as positive. The proof is on exactly the same lines as that of the Mean Value Theorem, but it starts with a more complicated function. If $b=a+h$, write:

$$F(x)=f(b)-f(x)-(b-x)f'(x)-\frac{(b-x)^2}{2!}f''(x)-\ldots-\frac{(b-x)^n}{n!}f^{(n)}(x).$$

So

$$F'(x)=-f'(x)+\left\{f'(x)-(b-x)f''(x)+(b-x)f''(x)-\frac{(b-x)^2}{2!}f'''(x)+\ldots\right.$$
$$\left.+\frac{(b-x)^{n-1}}{(n-1)!}f^{(n)}(x)-\frac{(b-x)^n}{n!}f^{(n+1)}(x)\right\}$$
$$=-\frac{(b-x)^n}{n!}f^{(n+1)}(x)\quad\text{(all other terms cancelling).}$$

Write: $$\phi(x)=F(x)-(b-x)^{n+1}\frac{F(a)}{(b-a)^{n+1}}.$$

So: $$\phi'(x)=F'(x)+(n+1)(b-x)^n\frac{F(a)}{(b-a)^{n+1}}$$
$$=(n+1)\frac{(b-x)^n}{(b-a)^{n+1}}\left\{F(a)-\frac{(b-a)^{n+1}}{(n+1)!}f^{(n+1)}(x)\right\}.$$

Now $\phi(a)=\phi(b)=0$ so that, by Rolle's Theorem, there is an $(a<\alpha<b)$ such that $\phi'(\alpha)=0$, i.e. such that

$$F(a)-\frac{(b-a)^{n+1}}{(n+1)!}f^{(n+1)}(\alpha)=0.$$

Put back $b=a+h$, i.e. $b-a=h$ and $\alpha=a+\theta h$ $(0<\theta<1)$.

$$F(a)=R_n\quad\text{where}\quad R_n=\frac{h^{n+1}}{(n+1)!}f^{(n+1)}(a+\theta h)$$

and $$F(a)=f(a+h)-f(a)-hf'(a)-\frac{h^2}{2!}f''(a)-\ldots-\frac{h^n}{n!}f^{(n)}(a).$$

Q.E.D.

From the point of view of the expansion of $f(x)$, the useful case of Taylor's Theorem is that with $a=0$, $h=x$:

$$f(x)=f(0)+f'(0)x+f''(0)\frac{x^2}{2!}+\ldots+f^{(n)}(0)\frac{x^n}{n!}+R_n\ldots\ldots\ldots\ldots(1)$$

where $$R_n = f^{(n+1)}(\theta x)\frac{x^{n+1}}{(n+1)!} \quad (0<\theta<1) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

R_n can be written in various ways; another one (11.9 Ex. 4) is the following:

$$R_n = f^{(n+1)}(\theta x)(1-\theta)^n\frac{x^{n+1}}{n!} \quad (0<\theta<1) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

The conditions for writing (1), to make Taylor's Theorem valid, are that $f(x)$, $f'(x)$, ... $f^{(n)}(x)$ are continuous over the interval $[0, x]$ and that $f^{(n+1)}(x)$ exists within the interval. To allow for various positive and negative values of x, we usually write (1) for $|x| \leqslant r$, where r is selected so that the function and its first n derivatives are continuous over the interval $[-r, r]$ about $x=0$ and so that the $(n+1)$th derivative exists within the interval.

We can now look ahead a little. Consider the sequence or series of terms

$$f(0)+f'(0)x+f''(0)\frac{x^2}{2!}+\ldots+f^{(n)}(0)\frac{x^n}{n!}+\ldots$$

where the '+' signs indicate that we propose to add the terms. This is called *Taylor's series*. What we have shown in (1) is that the sum of Taylor's series up to and including the term in x^n is an *approximation* to $f(x)$, *provided* that R_n given by (2) or (3) is small. In this case, we have an approximate representation of the function $f(x)$, whatever its form may be, as a polynomial of degree n in x. We would like to go further and say that $f(x)$ is exactly represented by the sum of Taylor's series continued indefinitely (the sum of the 'infinite series'). For this, we look for $R_n \to 0$ as $n \to \infty$. This is, in fact, the position we eventually attain. It all turns on what can be said about R_n, i.e. the *remainder* or difference between $f(x)$ and its polynomial approximation.

One further result can be added, the *Mean Value Theorem for integrals*:

THEOREM: *If $f(x)$ is continuous $a \leqslant x \leqslant b$, then α exists $(a \leqslant \alpha \leqslant b)$ so that:*

$$\int_a^b f(x)\,dx = f(\alpha)(b-a).$$

Things are simpler for integrals and all we need is the continuity of the function over the interval in question. The graphical interpretation is clear enough. In Fig. 11.1c, which has $f(x)$ positive in $[a, b]$, the area under the curve $y=f(x)$ between A and B is equal to the area of a rectangle $MCDN$ given by the height of the curve at some point P. The first area is $\int_a^b f(x)\,dx$; the rectangular area is $f(\alpha)(b-a)$. The proof comes directly from the definition of the integral as an area. Since $f(x)$ is continuous over $[a, b]$, it ranges over an interval $[c, d]$: $c \leqslant f(x) \leqslant d$. The area $\int_a^b f(x)\,dx$ is not greater than the rectangle $d(b-a)$, d being the LUB of $f(x)$, and it is not less than the rectangle $c(b-a)$. So: $c(b-a) \leqslant \int_a^b f(x)\,dx \leqslant d(b-a)$. But $f(x)$ attains every value between c and d. Therefore there is some α $(a \leqslant \alpha \leqslant b)$ so that $f(\alpha)(b-a)$ takes a specified value between $c(b-a)$ and $d(b-a)$, i.e. exists so that $f(\alpha)(b-a) = \int_a^b f(x)\,dx$. Q.E.D.

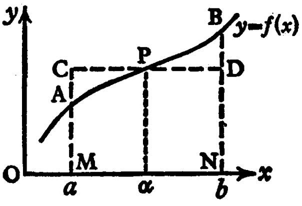


FIG. 11.1c

11.2. Maximum and minimum values. Suppose only that the function $y=f(x)$ is bounded on an interval $[a, b]$. Then there is a GLB c and a LUB d such that $c \leqslant f(x) \leqslant d$ in the interval $[a, b]$. It is not necessary, however, that $f(x)$ takes every value in the interval $[c, d]$; it happens to do so if $f(x)$ is continuous. The (extremely) hypothetical example, represented graphically in Fig. 11.2, illustrates the possibilities to be expected. The GLB c occurs at F where $x=\alpha$; the LUB d occurs at G

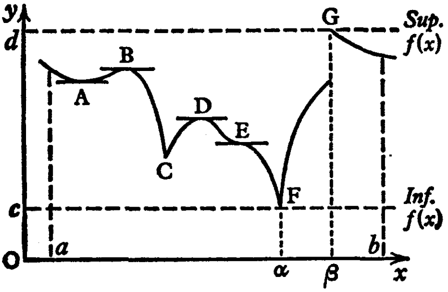


FIG. 11.2

where $x=\beta$. From the point of view of picking the smallest and largest values of $y=f(x)$ over the interval $[a, b]$, we may write:

$$\begin{aligned} &\operatorname{Inf} f(x)=c \quad \text{at} \quad x=\alpha \\ &\operatorname{Sup} f(x)=d \quad \text{at} \quad x=\beta \end{aligned}$$

where Inf stands for 'infemum' and Sup for 'supremum'.

A more important matter is the determination of *local* bounds, i.e. the smallest or largest values in a particular neighbourhood. This leads at once to the concept of (local) *maximum and minimum values*:

DEFINITION: *$f(x)$ has a* **maximum value** *Max $f(x)$ at $x=\alpha$ if there is a neighbourhood N of α such that $f(x)<f(\alpha)$ for all $x \in N$, $x \neq \alpha$; and a* **minimum value** *Min $f(x)$ if $f(x)>f(\alpha)$ under the same conditions.*

The possibilities are illustrated in Fig. 11.2; maximum values occur at B, D and G and minimum values at A, C and F. Being local concepts, several maximum (minimum) values can occur; and one maximum (e.g. at D) can be below some minimum (e.g. at A). In general, if $f(x)$ is not continuous, the determination of the maximum (minimum) values of $f(x)$ is a matter of direct use of the definition.

Concentrate now on a particular point $x=\alpha$ and *assume* that $f'(\alpha)$ exists. The sign taken by $f'(\alpha)$ is important in view of the interpretation of $f'(\alpha)$ as the rate of change of $f(x)$ at $x=\alpha$:

THEOREM: *If $f'(\alpha)>0$, then $f(x)$ is locally increasing at $x=\alpha$; if $f'(\alpha)<0$, then $f(x)$ is locally decreasing at $x=\alpha$.*

Proof: suppose $f'(\alpha)>0$ so that $\dfrac{f(\alpha+h)-f(\alpha)}{h}>0$ for sufficiently small h. Write $x=\alpha+h$ and there exists a neighbourhood N of α such that $f(x)-f(\alpha)$ has the same sign as $x-\alpha$ for all x of N. Hence $f'(\alpha)>0$ implies that $f(x)$ is below $f(\alpha)$ just to the left, and above $f(\alpha)$ just to the right, of $x=\alpha$. The opposite holds for $f'(\alpha)<0$. Q.E.D.

It follows immediately:

THEOREM: *A necessary condition for a maximum or minimum of $f(x)$ at α is $f'(\alpha)=0$, i.e. a maximum or minimum at α implies $f'(\alpha)=0$.*

For, if $f'(\alpha)>0$ or $f'(\alpha)<0$, the theorem above shows that there cannot be a maximum or minimum at $x=\alpha$. In Fig. 11.2, derivatives exist at the maxima B and D and at the minimum A; in all cases the derivative is zero (tangent horizontal). Derivatives do not exist at C,

F and G so that the theorem does not apply. Conversely, at the point E, the derivative is zero, but this is not a maximum or minimum value. The condition $f'(\alpha)=0$ applies only when the derivative exists and it is a necessary but not a sufficient condition.*

Various sufficient conditions can be found for a maximum (minimum) value. For example, if $f'(x)$ exists in a neighbourhood N of $x=\alpha$ and if $f'(x)>0$ for $x<\alpha$ in N, $f'(x)<0$ for $x>\alpha$ in N, then $f(\alpha)$ is a maximum value. Similarly, if the signs are reversed, i.e. $f'(x)<0$ to the left and $f'(x)>0$ to the right of $x=\alpha$, then $f(\alpha)$ is a minimum value.

Finally, for a function which has derivatives of all orders, a convenient set of *necessary and sufficient conditions* can be written:

THEOREM: *The necessary and sufficient conditions for $f(x)$ to have a maximum (minimum) at $x=\alpha$ are that the first non-zero derivative at $x=\alpha$ is of even order and that its sign is negative (positive).*

By Taylor's Theorem (11.1), if $f^n(\alpha)$ is the first non-zero derivative:

$$f(\alpha+h)=f(\alpha)+\frac{h^n}{n!}f^n(\alpha)+R_n$$

where $$R_n=\frac{h^{n+1}}{(n+1)!}f^{n+1}(\alpha+\theta h).$$

Now h can be taken so small that R_n is numerically less than $\frac{h^n}{n!}f^{(n)}(\alpha)$; this is because R_n involves the higher power h^{n+1}. Hence, for small h, the sign of $f(\alpha+h)-f(\alpha)$ is fixed by $\frac{h^n}{n!}f^n(\alpha)$. Suppose $f^{(n)}(\alpha)<0$ and n is even; then $\frac{h^n}{n!}f^n(\alpha)<0$ for all h, whether positive or negative. Hence, $f(\alpha+h)-f(\alpha)$ is negative for all small h, i.e. $f(\alpha)$ is a maximum. Similarly, if $f^n(\alpha)>0$ and n is even, then $f(\alpha)$ is a

* All local maximum and minimum values of $y=f(x)$ are sometimes described as *extreme values* of the function, i.e. an extreme value is either a maximum or a minimum. All values of $y=f(x)$ such that $f'(x)=0$ are sometimes called *stationary values* of the function, i.e. they comprise such maximum and minimum values and such points of inflexion as occur where $f'(x)$ exists and is zero. Of the seven points indicated in Fig. 11.2, all give extreme values except E, and all give stationary values except C, F and G. There are extreme values which are not stationary, and stationary values which are not extreme. However, for points where $f'(x)$ exists, the position is simpler: all extreme values are stationary. The converse is still not true, i.e. stationary values include points of inflexion (such as E) as well as maximum and minimum values.

minimum. The conditions are sufficient. Conversely, if $f(\alpha)$ is a maximum (minimum), then $f(\alpha+h)-f(\alpha)$ is negative (positive) for all sufficiently small h. Hence $\frac{h^n}{n!}f^n(\alpha)$ is negative (positive) for all sufficiently small h and the stated conditions follow. The conditions are necessary. Q.E.D.

The result is most suitable for use in practice. If $f(x)$ has derivatives of all required orders at $x=\alpha$, we proceed as follows. First, check that $f'(\alpha)=0$. Second, write $f''(\alpha)$ and see if it is non-zero; $f''(\alpha)<0$ implies that $f(\alpha)$ is a maximum and $f''(\alpha)>0$ that $f(\alpha)$ is a minimum. If $f''(\alpha)=0$, write $f'''(\alpha)$; a non-zero value here means that $f(\alpha)$ is neither a maximum nor a minimum. This is a case of a *point of inflexion*, as at the point E of Fig. 11.2.* If $f'''(\alpha)=0$, write $f''''(\alpha)$; if this is not zero, its sign determines whether $f(\alpha)$ is a maximum (negative sign) or a minimum (positive sign). The process goes on until a non-zero derivative is reached. Two examples:

(i) $y=x^3-3x^2+5$ for which $Dy=3x^2-6x=3x(x-2)$ and $D^2y=6x-6=6(x-1)$.

Hence $Dy=0$ at $x=0$ and at $x=2$.

At $x=0$, $D^2y=-6<0$ i.e. $y=5$ at $x=0$ is a maximum

At $x=2$, $D^2y=6>0$ i.e. $y=1$ at $x=2$ is a minimum.

(ii) $y=x^3-3x^2+3x+5$ for which $Dy=3x^2-6x+3=3(x-1)^2$
$D^2y=6x-6=6(x-1)$ and
$D^3y=6$.

Hence, at $x=1$: $Dy=D^2y=0$, $D^3y=6\neq0$. This is a point of inflexion. Since $Dy=0$ only at $x=1$, there are no maximum or minimum values.

11.3. Convergence of series. A sequence of terms u_n (for $n=1, 2, 3, \ldots$) is a countably infinite set. Suppose we wish to add them. From this point of view, the terms are an *infinite series*, written $u_1+u_2+u_3+\ldots$, or more shortly as Σu_n. The sum of n terms can always be written:

$$S_n=\sum_{r=1}^{n} u_r=u_1+u_2+u_3+\ldots+u_n.$$

* A point of inflexion is one at which the curve turns over the tangent; it can occur for upward or downward sloping, as well as for horizontal, tangents.

The question is whether any meaning can be attached to the sum of the whole infinite series. The fact that this is a real question, and probably not an easy one to answer, is clear from the following argument:

Write $\quad a=1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\ldots\quad$ and $\quad b=1+\frac{1}{3}+\frac{1}{5}+\frac{1}{7}+\ldots$

Then

$$a=(1+\tfrac{1}{3}+\tfrac{1}{5}+\ldots)+(\tfrac{1}{2}+\tfrac{1}{4}+\tfrac{1}{6}+\ldots)=b+\tfrac{1}{2}(1+\tfrac{1}{2}+\tfrac{1}{3}+\ldots)=b+\tfrac{1}{2}a.$$

Hence $b=\frac{1}{2}a$. Now write $c=1-\frac{1}{2}+\frac{1}{3}-\frac{1}{4}+\ldots$ so that

$$c=(1+\tfrac{1}{3}+\tfrac{1}{5}+\ldots)-(\tfrac{1}{2}+\tfrac{1}{4}+\tfrac{1}{6}+\ldots)=b-\tfrac{1}{2}a=0.$$

But $\quad c=(1-\frac{1}{2})+(\frac{1}{3}-\frac{1}{4})+(\frac{1}{5}-\frac{1}{6})+\ldots>0.$

There is obviously something amiss. There is a fallacy, and it is in the first line, in the all-too-facile assumption that these two infinite series do have sums a and b. It will be seen that this assumption is wrong.

The sum $S_n=\sum_{r=1}^{n} u_r$ is itself a sequence for $n=1, 2, 3, \ldots$. The ideas of 9.4 and 9.5 on limits of sequences are of immediate application and we may ask whether S_n has a limit as $n \to \infty$. If so, the series is *convergent*:

DEFINITION: *If* $S_n=\sum_{r=1}^{n} u_r \to S$ *as* $n\to\infty$, *the infinite series* Σu_n *is* **convergent** *and its sum is* S.

In other cases, the infinite series is not convergent and there is no sum. The properties of limits provide two simple results: if Σu_n converges to S and if k is a constant, then Σv_n where $v_n=ku_n$ converges to kS; if Σu_n converges to S_1 and Σv_n converges to S_2, then Σw_n where $w_n=u_n+v_n$ converges to S_1+S_2.

The theorem of 9.5 on conditions for a limit translates into an appropriate form for convergence of series:

THEOREM: Σu_n *is convergent if and only if, given* ϵ, *there is an integer* N *such that* $|S_m-S_n|\leqslant\epsilon$ *for all* m *and* n *both greater than* N.

The theorem of 9.5 states that, given ϵ (and so given $\frac{1}{2}\epsilon$), there is an integer N such that $S-\frac{1}{2}\epsilon\leqslant S_p\leqslant S+\frac{1}{2}\epsilon$ for all $p>N$, if S_n is to tend to S as $n\to\infty$. Hence, if both m and n are greater than N:

$$S-\tfrac{1}{2}\epsilon\leqslant S_m\leqslant S+\tfrac{1}{2}\epsilon \quad\text{and}\quad S-\tfrac{1}{2}\epsilon\leqslant S_n\leqslant S+\tfrac{1}{2}\epsilon$$

i.e. S_m and S_n can differ by ϵ at most. Q.E.D.

As a *necessary condition* for convergence, $u_n \to 0$ as $n \to \infty$. This follows from the theorem just proved. We have: $u_n = S_n - S_{n-1}$. Suppose Σu_n is convergent, so that, given ϵ, we can select an integer (say) $N-1$ such that $|S_n - S_{n-1}| \leqslant \epsilon$ for n and $n-1$ both greater than $N-1$. Hence $|u_n| \leqslant \epsilon$ for all $n > N$, i.e. $u_n \to 0$ as $n \to \infty$. The condition, however, is by no means a sufficient one; we find many examples of series which have $u_n \to 0$ as $n \to \infty$ but which are not convergent.

The various possibilities can be illustrated by a number of examples:

	A: Alternating Series		*B: Positive Series*	
(i)	$1 - \frac{3}{4} + \frac{9}{16} - \frac{27}{64} + \dots$	*C*	$1 + \frac{3}{4} + \frac{9}{16} + \frac{27}{64} + \dots$	*C*
(ii)	$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots$	*C*	$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots$	*NC*
(iii)	$1 - 1 + 1 - 1 + \dots$	*NC*	$1 + 1 + 1 + 1 + \dots$	*NC*
(iv)	$1 - \frac{3}{2} + \frac{9}{4} - \frac{27}{8} + \dots$	*NC*	$1 + \frac{3}{2} + \frac{9}{4} + \frac{27}{8} + \dots$	*NC*

Four pairs of infinite series are shown; the second of each pair is a series of positive terms, the first the same series but with terms alternating in sign. The pairs are arranged in an obvious order; the terms of (i) decrease rapidly, those of (ii) less rapidly, while (iii) and (iv) have terms which do not decrease. By the necessary condition, though (i) and (ii) may be convergent, (iii) and (iv) are certainly not. In the table, *C* indicates a convergent series and *NC* one which is not convergent, as established by the following analysis:

(i) These are geometric progressions of the form $1 + r + r^2 + r^3 + \dots$ with $S_n = \dfrac{1-r^n}{1-r}$ $(r \neq 1)$. Here $r = \pm\frac{3}{4}$ so that:

$$\text{A:} \quad S_n = \frac{1-(-\frac{3}{4})^n}{1-(-\frac{3}{4})} = \tfrac{4}{7}\{1-(-\tfrac{3}{4})^n\} \to \tfrac{4}{7} \text{ as } n \to \infty$$

$$\text{B:} \quad S_n = \frac{1-(\frac{3}{4})^n}{1-\frac{3}{4}} = 4\{1-(\tfrac{3}{4})^n\} \to 4 \text{ as } n \to \infty.$$

Both series are convergent, A with sum $\frac{4}{7}$ and B with sum 4.

(ii) There is no simple or obvious expression for S_n, for all integral n, and indirect methods need to be adopted in determining the convergence or otherwise of the series. Equally, in the absence of an explicit form for S_n, the sum S of the infinite series (if convergent) cannot be found directly.

$$\text{A:}\quad S_{2n+1}=S_{2n-1}-\frac{1}{2n}+\frac{1}{2n+1}<S_{2n-1}\qquad\text{since }\frac{1}{2n}>\frac{1}{2n+1}$$

$$S_{2n}=S_{2n-2}+\frac{1}{2n-1}-\frac{1}{2n}>S_{2n-2}\qquad\text{since }\frac{1}{2n-1}>\frac{1}{2n}$$

for any positive integer n. Hence, the odd sums S_1, S_3, S_5, ... form a decreasing sequence, all of them being less than S_1; the even sums S_2, S_4, S_6, ... form an increasing sequence, all greater than S_2. Further: $S_{2n+1}-S_{2n}=\dfrac{1}{2n+1}$ which can be made as small as we please for sufficiently large n. Hence the decreasing sequence of odd sums tends to the same limit S as the increasing sequence of even sums and $S_2<S<S_1$, i.e. $\frac{1}{2}<S<1$. The series is convergent with a sum which we cannot yet specify, except that it lies between $\frac{1}{2}$ and 1.

$$\text{B:}\quad 1+\tfrac{1}{2}+\tfrac{1}{3}+\tfrac{1}{4}+\ldots=1+\tfrac{1}{2}+(\tfrac{1}{3}+\tfrac{1}{4})+(\tfrac{1}{5}+\tfrac{1}{6}+\tfrac{1}{7}+\tfrac{1}{8})+\ldots$$
$$>1+\tfrac{1}{2}+\tfrac{1}{2}+\tfrac{1}{2}+\ldots$$

since $\frac{1}{3}>\frac{1}{4}$, i.e. $\frac{1}{3}+\frac{1}{4}>\frac{2}{4}=\frac{1}{2}$, and similarly for the other groups. Hence, if enough terms of the series are taken to make up n groups, the sum is greater than $1+\dfrac{n-1}{2}=\dfrac{n+1}{2}\to\infty$ as $n\to\infty$. The series is not convergent; the sum S_n increases steadily and indefinitely with n.

(iii) These series are not convergent since the general term does not tend to zero. Since expressions for S_n can be written, we can say something more.

A: $S_n=1$ (n odd) and $S_n=0$ (n even)

B: $S_n=n\to\infty$ as $n\to\infty$.

Hence series A oscillates between finite limits (0 and 1) but series B has a sum which increases steadily and indefinitely with n.

(iv) These series are also not convergent; indeed $|u_n|\to\infty$ as $n\to\infty$ for each. They are geometric progressions with $r=\pm\frac{3}{2}$; the sums S_n are as follows.

$$\text{A:}\quad S_n=\frac{1-(-\frac{3}{2})^n}{1-(-\frac{3}{2})}=\tfrac{2}{5}\{1-(-\tfrac{3}{2})^n\}$$

and this oscillates with increasing amplitude as n increases.

$$\text{B:}\quad S_n=\frac{(\frac{3}{2})^n-1}{\frac{3}{2}-1}=2\{(\tfrac{3}{2})^n-1\}\to\infty\text{ as }n\to\infty.$$

Hence the series A oscillates and series B does not; in both cases, the absolute value of S_n increases indefinitely with n.

Some general conclusions can be drawn. First, there are only two possibilities for a series of positive terms: *either* the series is convergent with S_n increasing steadily to S, *or* the series is not convergent with S_n increasing steadily and indefinitely. Second, there is a greater variety of possibilities for a series with mixed positive and negative terms, including cases of (finite or indefinite) oscillation of S_n. Third, the series of alternating terms may often be the same (simply as regards convergence or otherwise) as the series of positive terms. But there is the interesting marginal case, illustrated by (ii), where the alternating series is convergent and the positive series not convergent.

11.4. Series of positive terms. The series Σu_n, where $u_n > 0$ all n, has a sum S_n which is positive and increasing with n. This simplifies the situation considerably:

either S_n is bounded above, increasing to a limit S, with Σu_n convergent to S;

or S_n is not bounded above, increasing without limit, with Σu_n not convergent.

For series where no simple expression for S_n is available, we require tests which distinguish between the alternatives. The simplest of such tests turns on:

Comparison Theorem: *If $u_n \leqslant kv_n$ where Σv_n is convergent and k a positive constant, then Σu_n is convergent; if $u_n \geqslant kv_n$ where Σv_n is not convergent and k a positive constant, then Σu_n is not convergent.*

The proof is simple. In the first case, if $S_n = \sum_{r=1}^{n} u_r$ and $S_n' = \sum_{r=1}^{n} v_n$, then $S_n \leqslant kS_n'$. But S_n' has a limit ($n \to \infty$) which is an upper bound of S_n'; hence S_n is bounded above and Σu_n is convergent. Similarly for the other case.

An obvious series for comparison is the geometric series:

$$1 + r + r^2 + r^3 + \ldots$$

where r is positive. This is convergent to $\frac{1}{1-r}$ if $r < 1$ and not convergent if $r \geqslant 1$. The following tests* then follow:

CAUCHY'S TEST: *If* $u_n \leqslant kr^{n-1}$ *(all* n*), where* $0 < r < 1$ *and* k *a positive constant, then* Σu_n *is convergent to sum* $S \leqslant \frac{k}{1-r}$.

D'ALEMBERT'S TEST: *If* $\frac{u_n}{u_{n-1}} \leqslant r$ *(all* $n > 1$*), where* $0 < r < 1$*, then* Σu_n *is convergent to sum* $S \leqslant \frac{u_1}{1-r}$.

The proof of the first is a direct application of the Comparison Theorem. The second is reduced to the first by writing:

$$u_n = \frac{u_n}{u_{n-1}} \frac{u_{n-1}}{u_{n-2}} \cdots \frac{u_2}{u_1} u_1 \leqslant r^{n-1} u_1.$$

Since a finite number of terms can always be added to or removed from a series without affecting its convergence, the tests apply equally well if the conditions hold for sufficiently large n. (The sum S, however, is changed.) Hence there is a more practical form of the tests using limits:

If $\sqrt[n]{u_n}$ or $\frac{u_n}{u_{n-1}}$ tends to a limit less than 1, Σu_n is convergent.

Proof: suppose $\sqrt[n]{u_n} \to L$ as $n \to \infty$ $(L < 1)$. Given ϵ, there is an integer N such that $L - \epsilon \leqslant \sqrt[n]{u_n} \leqslant L + \epsilon$ for all $n > N$. Write $L + \epsilon = r$ and choose ϵ so that $r < 1$ (as well as $L < 1$). Then $\sqrt[n]{u_n} \leqslant r$, i.e. $u_n \leqslant r^n$, for all $n > N$, which is Cauchy's Test (with $k = r$). The other case follows similarly.

It can be noticed, in passing, that the question of the convergence of an infinite series is matched by a similar question of the convergence of an infinite integral. The integral $\int_a^k f(x)\,dx$ can always be written if $f(x)$ is continuous for $x \geqslant a$. It remains to investigate whether $\int_a^k f(x)\,dx \to L$ as $k \to \infty$. If so, the infinite integral exists: $\int_a^\infty f(x)\,dx = L$ (see 10.9 Ex. 24). A further test of the convergence of an infinite series of positive terms can be derived from consideration of an infinite integral (11.9 Ex. 19).

* The first test was given by Cauchy (1789–1857) in the particular case where $k = 1$; the second test was given by d'Alembert (1717–83).

11.5. Absolute and conditional convergence. When an infinite series consists of mixed positive and negative terms, the problem of testing for convergence (in the absence of a simple expression for S_n) is complicated by the greater variety of possibilities. A series may fail to be convergent for several reasons: S_n may increase steadily and indefinitely through positive values, or decrease steadily and indefinitely through negative values, or oscillate between finite limits or between indefinitely increasing values.* The best procedure in practice is to start with the corresponding series of positive terms, i.e. to get at the convergence of Σu_n by first examining $\Sigma \mid u_n \mid$.

Suppose that $\Sigma \mid u_n \mid$ is convergent. Then the series Σv_n consisting of the positive terms of Σu_n is convergent. Equally, the series Σw_n which consists of the negative terms of Σu_n, each with sign changed, is convergent. Here, $\Sigma \mid u_n \mid = \Sigma v_n + \Sigma w_n$ and $\Sigma u_n = \Sigma v_n - \Sigma w_n$. Hence, if $\Sigma \mid u_n \mid$ is convergent, so is Σu_n and it is then said to be 'absolutely convergent':

DEFINITION: *The series Σu_n is* **absolutely convergent** *if $\Sigma \mid u_n \mid$ is convergent; Σu_n is necessarily convergent if it is absolutely convergent. The series Σu_n is* **conditionally convergent** *if it is convergent but $\Sigma \mid u_n \mid$ is not convergent.*

An illustration of absolute convergence is a geometric series

$$1 + r + r^2 + r^3 + \dots$$

with $\mid r \mid < 1$. The fact that, if $\Sigma \mid u_n \mid$ is not convergent, then Σu_n *may* be convergent, is illustrated by example (ii) of 11.3. The series $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots$ is conditionally convergent; it is convergent while $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots$ is not.

For testing series of mixed terms, if not absolutely convergent, we have very little to go on; indeed, there is only one result of general assistance:

THEOREM: *If u_n is positive and decreases* **steadily** *to zero as $n \to \infty$, then the alternating series $u_1 - u_2 + u_3 - u_4 + \dots$ is convergent.*

The proof follows on exactly the lines used in the particular case of example (ii) A of 11.3. The sum of the series lies between $u_1 - u_2$ and u_1.

* The negative term 'non-convergent' covers all these possibilities, but more positive labels are variously attached by different writers. Some describe all non-convergent series as 'divergent'; others distinguish between 'divergent' and 'oscillatory' series.

Two further examples illustrate:

(i) The series $1+1+\frac{1}{2!}+\frac{1}{3!}+\ldots$ is convergent by d'Alembert's test, since

$$u_n=\frac{1}{(n-1)!} \quad \text{and} \quad \frac{u_n}{u_{n-1}}=\frac{(n-2)!}{(n-1)!}=\frac{1}{n-1}<1 \quad \text{for } n>2$$

The series $1-1+\frac{1}{2!}-\frac{1}{3!}+\ldots$ is absolutely convergent.

(ii) The series $1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\ldots$ is convergent by the theorem above. But:

$$\begin{aligned}1+\tfrac{1}{3}+\tfrac{1}{5}+\tfrac{1}{7}+\ldots &= 1+\tfrac{1}{3}+(\tfrac{1}{5}+\tfrac{1}{7})+(\tfrac{1}{9}+\tfrac{1}{11}+\tfrac{1}{13}+\tfrac{1}{15})+\ldots\\ &> 1+\tfrac{1}{4}+\tfrac{1}{4}+\tfrac{1}{4}+\ldots\end{aligned}$$

since $\frac{1}{3}>\frac{1}{4}$, $\frac{1}{5}+\frac{1}{7}>\frac{1}{8}+\frac{1}{8}=\frac{2}{8}=\frac{1}{4}$, and so on. Hence, if enough terms of the series are taken to make up n groups, the sum exceeds

$$1+\frac{n-1}{4}=\frac{n+3}{4}$$

i.e. the series is not convergent. Consequently, $1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\ldots$ is only conditionally convergent.

A convergent series can be used, by taking enough terms, to give an approximation to its sum. For example, if the sum of the series $1+1+\frac{1}{2!}+\frac{1}{3!}+\ldots$ is denoted by the constant e, a rational approximation to e (which is irrational) is:

$$\begin{aligned}e &= 1+1+\tfrac{1}{2}+\tfrac{1}{6}+\tfrac{1}{24}+\tfrac{1}{120}+\tfrac{1}{720}+\tfrac{1}{5040}+\ldots\\ &=2{\cdot}5\\ &+0{\cdot}1667\\ &+0{\cdot}0417\\ &+0{\cdot}0083\\ &+0{\cdot}0014\\ &+0{\cdot}0002+\ldots\\ &=2{\cdot}718 \quad \text{to three decimal places.}\end{aligned}$$

Seven terms of the series are required for this approximation.

11.6. Power series. The development of 11.1 suggests that Taylor's Series may give an approximation to a given function $f(x)$ as a polynomial in ascending powers of x, and an exact representation of

$f(x)$ as the sum of an infinite series. The remainder R_n in Taylor's Theorem is critical; it must be small for the approximation and it must tend to zero as $n \to \infty$ for the exact representation. This is the matter now to be pursued, i.e. the *expansion* of a function $f(x)$ in ascending powers of x.

If x is a real variable and $a_0, a_1, a_2, \ldots a_n, \ldots$ a set of constants, then $\Sigma a_n x^n = a_0 + a_1 x + a_2 x^2 + \ldots$ is a *power series.* Given a value of x, the series may be tested for convergence in the ways described above. It may turn out to be convergent for no x $(x \neq 0)$ or for all x; examples of both cases are given later. In general, however, it is convergent for some and not for other values of x. A most convenient result can be established; it is that a power series is convergent over an interval (open or closed) around $x=0$ and absolutely convergent within the interval. It is just not possible for the series to be convergent for disconnected values of x. This result is reached in two stages.

THEOREM: *If $\Sigma a_n x^n$ is convergent at $x=\alpha$, then it is absolutely convergent for $-r<x<r$, where $r=|\alpha|$.*

Proof: since $\Sigma a_n \alpha^n$ is convergent, it is necessary that $a_n \alpha^n \to 0$ as $n \to \infty$, i.e. that $a_n \alpha^n$ is bounded: $|a_n \alpha^n|<k$ for some positive k and all n. Hence:

$$|a_n x^n| = |a_n \alpha^n| \, |x/\alpha|^n < k \, |x/\alpha|^n$$

so that, by Cauchy's Test (11.4), $\Sigma \, |a_n x^n|$ is convergent if $|x/\alpha|<1$, i.e. $\Sigma a_n x^n$ is absolutely convergent if $|x|<|\alpha|=r$. Q.E.D.

THEOREM: *Given a power series $\Sigma a_n x^n$, there are three possibilities:*

(i) *it is convergent for no non-zero x*

or (ii) *it is absolutely convergent for all x*

or (iii) *it is absolutely convergent for $|x|<r$, and not convergent for $|x|>r$, where r is some positive constant.*

Notice that, in case (iii), nothing is said about the convergence of the series for $x= \pm r$; this is a matter left entirely open. The proof is: Partition the set of all real numbers $x \geqslant 0$ into two subsets A and B, where A consists of x such that $\Sigma a_n x^n$ is convergent and B consists of x such that $\Sigma a_n x^n$ is not convergent. A is not empty since it contains $x=0$. B consists only of positive x but it may be empty. If B is not empty, then $\alpha < \beta$ for all $\alpha \in A$ and $\beta \in B$. This follows from two applications of the theorem just established. If $\alpha \in A$, $\Sigma a_n \alpha^n$ is convergent and $\Sigma a_n x^n$ is convergent for all $x<\alpha$, i.e. there is no $\beta \in B$ which is less than α. Conversely, if $\beta \in B$, $\Sigma a_n \beta^n$ is not convergent and $\Sigma a_n x^n$ cannot be convergent at any $x=\alpha>\beta$; otherwise it would need to be convergent also at β. Hence there is no $\alpha \in A$ which is greater than β. Finally $\alpha=\beta$ is ruled out since A and B are disjoint sets. Hence $\alpha<\beta$ always, i.e. all elements of A are less than all elements of B. If r is the LUB of set A, the dividing value between A and B is r, belonging either to A or to B. Similarly, the set of real $x \leqslant 0$ is partitioned into two subsets A' and B' for convergence and non-convergence of $\Sigma a_n x^n$ respectively. All elements of A' are less in absolute value than all elements of B' and there is a dividing value $-r'$. The preceding theorem then shows that the interval of convergence must be symmetrical (i.e. $r=r'$) and that, within the interval, $\Sigma a_n x^n$ is not only convergent but also absolutely convergent. Pulling the cases together, we find three possibilities: A and A' contain only $x=0$ which is case (i); or B and B' are both empty which is case (ii); or there is a positive r such that the series is absolutely convergent for $-r<x<r$ and not convergent for $x>r$ and $x<-r$. Q.E.D.

Hence, in handling $\Sigma a_n x^n$, we look for the marginal values $x=\pm r$. The series is absolutely convergent within the interval $[-r, r]$ and not convergent outside it. At $x=\pm r$, the series may or may not be convergent. When found, r is called the *radius of convergence* of the power series. Illustrative examples follow, arranged under the headings (i), (ii) and (iii) of the three cases:

(i) $1+x+2!x^2+3!x^3+\ldots+n!x^n+\ldots$ is convergent for no non-zero x.

Here: $u_n=n!x^n=nx\times(n-1)!x^{n-1}=nx\times u_{n-1}$ if the series is $\sum\limits_{n=0}^{\infty} u_n$

i.e. $$|u_n|=n|x||u_{n-1}|.$$

No matter what non-zero x is taken, there is n sufficiently large for $n|x|>1$, i.e. for $|u_n|>|u_{n-1}|$. Hence, u_n cannot tend to zero as $n\to\infty$ and the series is not convergent for any non-zero x.

(ii) $1+x+\dfrac{x^2}{2!}+\dfrac{x^3}{3!}+\ldots+\dfrac{x^n}{n!}+\ldots$ is absolutely convergent for all x.

Here: $$u_n=\frac{x^n}{n!} \quad\text{and}\quad \left|\frac{u_n}{u_{n-1}}\right|=\frac{|x|}{n}\to 0 \text{ as } n\to\infty.$$

By d'Alembert's Test (11.4), $\Sigma \mid u_n \mid$ is convergent, i.e. Σu_n is absolutely convergent, all x. The same result holds for other series obtained by taking some (but not all) terms of the series, and by varying their signs. For example:

$$1-\frac{x^2}{2!}+\frac{x^4}{4!}-\frac{x^6}{6!}+\ldots; \quad 1+\frac{x^2}{2!}+\frac{x^4}{4!}+\frac{x^6}{6!}+\ldots;$$

$$x-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+\ldots; \quad x+\frac{x^3}{3!}+\frac{x^5}{5!}+\frac{x^7}{7!}+\ldots$$

are each absolutely convergent for all x.

(iii) (*a*) $x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\ldots+(-1)^{n-1}\frac{x^n}{n}+\ldots$ is absolutely convergent, $-1<x<1$. The series is convergent at $x=1$, not convergent at $x=-1$. The conditional convergence of the series when $x=1$ follows from example (ii) of 11.3. Hence, by the first theorem above, the power series is absolutely convergent for $-1<x<1$ and not convergent outside this interval.

(*b*) $x-\frac{x^3}{3}+\frac{x^5}{5}-\frac{x^7}{7}+\ldots+(-1)^{n-1}\frac{x^{2n-1}}{2n-1}+\ldots$ is absolutely convergent, $-1<x<1$. It is conditionally convergent for $x=\pm 1$. When $x=1$, the series is $1-\frac{1}{3}+\frac{1}{5}-\frac{1}{7}+\ldots$ which is conditionally convergent by example (ii) of 11.5. The same series is obtained, with all signs reversed, when $x=-1$. The absolute convergence of the power series for $-1<x<1$ again follows by the first theorem above.

11.7. Expansions of functions. A function $f(x)$ has derivatives of all orders for all values of x (at least in a certain interval around $x=0$). Then, by Taylor's Theorem of 11.1:

$$f(x)=f(0)+f'(0)x+f''(0)\frac{x^2}{2!}+\ldots+f^{(n)}(0)\frac{x^n}{n!}+R_n\ldots\ldots\ldots(1)$$

where $$R_n=f^{(n+1)}(\theta x)\frac{x^{n+1}}{(n+1)!} \quad \text{or} \quad f^{(n+1)}(\theta x)(1-\theta)^n\frac{x^{n+1}}{n!} \quad \ldots\ldots(2)$$

for some real θ between 0 and 1.

The next step is the difficult one: to find what values of x have $R_n \to 0$ as $n \to \infty$. By the main theorem for power series (11.6), we expect either no x (except $x=0$), or all x, or x within a certain interval

$[-r, r]$, where r is the radius of convergence of the power series written by continuing (1) indefinitely:

$$f(x)=f(0)+f'(0)x+f''(0)\frac{x^2}{2!}+\ldots+f^{(n)}(0)\frac{x^n}{n!}+\ldots \qquad \ldots\ldots\ldots(3)$$

The first two cases can be included in the third by allowing $r=0$ and $r\to\infty$. The difficulty remains; it is to find the appropriate r. If this can be done:

THEOREM: *If $f(x)$ has derivatives of all orders and if R_n given by* (2) *tends to zero as $n\to\infty$ for all x within the interval $-r<x<r$, then $f(x)$ can be expanded as the power series* (3), *absolutely convergent for $-r<x<r$.*

There is uncertainty whether or not the expansion holds also for $x=\pm r$, a point which needs examination in each case.

To illustrate the problem of examining the limit of R_n and hence of fixing the radius of convergence, consider the expansion of the power function $(1+x)^m$, where x is a real variable and m is a given rational exponent. This is one of the most frequently used of all expansions, a very important one to establish. Two particular cases can be conveniently taken before the general case.

Case: $m=-1$, i.e. the expansion of $(1+x)^{-1}$.

The geometric series $1-x+x^2-x^3+\ldots$ has $S_n=\dfrac{1-(-x)^n}{1+x}\to\dfrac{1}{1+x}$ as $n\to\infty$, provided that $|x|<1$. Hence we have the expansion:

$$(1+x)^{-1}=\frac{1}{1+x}=1-x+x^2-x^3+\ldots \quad \text{for} \quad -1<x<1 \ \ldots\ldots\ldots(4)$$

It can be checked that the coefficients are those written in (3). We have

$$D\left(\frac{1}{1+x}\right)=-\frac{1}{(1+x)^2}; \quad D^2\left(\frac{1}{1+x}\right)=D\left\{-\frac{1}{(1+x)^2}\right\}=\frac{2}{(1+x)^3}; \quad \ldots$$

Generally: $$D^n\left(\frac{1}{1+x}\right)=(-1)^n\frac{n!}{(1+x)^{n+1}}$$

and $$\left[D^n\left(\frac{1}{1+x}\right)\right]_{x=0}=(-1)^n n!$$

So (3) gives $\dfrac{1}{1+x}=1-x+x^2-x^3+\ldots$; (4) simply adds the fact that the radius of convergence is $r=1$, that the series is absolutely convergent for $|x|<1$.

Case: $m=n$ (positive integer), i.e. the expansion of $(1+x)^n$. Here the expansion is a finite polynomial of degree n, the infinite series of (3) terminating with the term in x^n. The polynomial is given by the *Binomial Theorem* of elementary algebra:

$$(1+x)^n=\sum_{r=0}^{n}\binom{n}{r}x^r=1+\binom{n}{1}x+\binom{n}{2}x^2+\ldots+\binom{n}{n-1}x^{n-1}+x^n \quad \ldots\ldots(5)$$

where $$\binom{n}{r}=\frac{n!}{r!\,(n-r)!} \qquad r=1,\,2,\,3,\,\ldots\,n-1$$

with the convention that $\binom{n}{0}=\binom{n}{n}=1$ (see 1.7 above). For particular values of the integers n and r, the values of $\binom{n}{r}$ can be got from the formula or, more easily, from Pascal's Triangle (5.8 above). For $n=2$ and $n=3$, (5) gives:

$$(1+x)^2=1+2x+x^2 \quad \text{and} \quad (1+x)^3=1+3x+3x^2+x^3$$

as can be checked by squaring and cubing $1+x$ directly. It can again be checked that the coefficients of (5) agree with those written in (3). For:

$$\begin{aligned} D^r(1+x)^n &= n(n-1)(n-2)\ldots(n-r+1)(1+x)^{n-r} & (r\leqslant n)\\ &=0 & (r>n)\end{aligned}$$

Hence, the last term in (3) is x^n, all later ones being zero. For $r\leqslant n$, the term in x^r is: $n(n-1)(n-2)\ldots(n-r+1)\dfrac{x^r}{r!}=\dfrac{n!}{r!\,(n-r)!}x^r=\dbinom{n}{r}x^r$ as given in (5)

General Case, the expansion of $(1+x)^m$, m any rational. The general derivative is:

$$D^n(1+x)^m=m(m-1)(m-2)\ldots(m-n+1)(1+x)^{m-n}$$

and so: $$\Big[D^n(1+x)^m\Big]_{x=0}=m(m-1)(m-2)\ldots(m-n+1).$$

Then (1) and (2) give:

$$(1+x)^m=1+mx+\frac{m(m-1)}{2!}x^2+\frac{m(m-1)(m-2)}{3!}x^3+\ldots$$
$$+\frac{m(m-1)(m-2)\ldots(m-n+1)}{n!}x^n+R_n$$

where

$$R_n = m(m-1)(m-2)\ldots(m-n)(1+\theta x)^{m-n-1}(1-\theta)^n \frac{x^{n+1}}{n!} \quad (0<\theta<1).$$

Some awkward algebra is needed to determine whether (and for what x) $R_n \to 0$ as $n \to \infty$. Write $R_n = \phi(n)\psi(n)$ where:

$$\phi(n) = \frac{m(m-1)(m-2)\ldots(m-n)}{n!} x^{n+1}; \quad \psi(n) = \frac{(1-\theta)^n}{(1+\theta x)^{n-m+1}}.$$

Then
$$\phi(n) = \frac{m-n}{n} x\phi(n-1)$$

i.e.
$$\left|\frac{\phi(n)}{\phi(n-1)}\right| = \left|\frac{m}{n} - 1\right| |x| \to |x| \quad \text{as } n \to \infty.$$

This limit is less than one, and so $|\phi(n)| < |\phi(n-1)|$ for sufficiently large n, provided that $|x|<1$. In this case, $\phi(n)$ is bounded for all n. Further $\psi(n) = \left(\frac{1-\theta}{1+\theta x}\right)^n (1+\theta x)^{m-1} \to 0$ as $n \to \infty$ if $|x|<1$. This is so since, given $0<\theta<1$ and $|x|<1$, $0<\left(\frac{1-\theta}{1+\theta x}\right)^n<1$ and so $\left(\frac{1-\theta}{1+\theta x}\right)^n \to 0$ as $n \to \infty$; $(1+\theta x)^{m-1}$ is not affected, being independent of n. Hence, $R_n = \phi(n)\psi(n) \to 0$ as $n \to \infty$ if $|x|<1$. We have what we need and we can write:

$$\left.\begin{aligned}(1+x)^m &= 1 + mx + \frac{m(m-1)}{2!}x^2 + \frac{m(m-1)(m-2)}{3!}x^3 + \ldots \\ &+ \frac{m(m-1)(m-2)\ldots(m-n+1)}{n!}x^n + \ldots \quad \text{for } -1<x<1\end{aligned}\right\} \quad \ldots\ldots(6)$$

The expansion (6) is called the *Binomial Series*. When m is a positive integer, (6) reduces to the finite series (5) of the Binomial Theorem. For other rational m, the series (6) does not terminate. Some examples:

(i) $m = -1$: general term $= \frac{(-1)(-2)(-3)\ldots(-n)}{n!}x^n = (-1)^n x^n$

and the expansion is:

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + \ldots \quad (-1<x<1)$$

which is the particular case (4) already written.

(ii) $m=\frac{1}{2}$: general term $=\dfrac{\frac{1}{2}(-\frac{1}{2})(-\frac{3}{2})\ldots(-n+\frac{3}{2})}{n!}x^n$

$$=(-1)^{n-1}\frac{1\,.\,3\,.\,5\ldots(2n-3)}{2\,.\,4\,.\,6\ldots 2n}x^n \quad (n>1)$$

and the expansion is:

$$\sqrt{1+x}=1+\tfrac{1}{2}x-\frac{1}{2\,.\,4}x^2+\frac{1\,.\,3}{2\,.\,4\,.\,6}x^3-\frac{1\,.\,3\,.\,5}{2\,.\,4\,.\,6\,.\,8}x^4+\ldots$$

$$=1+\tfrac{1}{2}x-\tfrac{1}{8}x^2+\tfrac{1}{16}x^3-\tfrac{5}{128}x^4+\ldots \quad (-1<x<1).$$

(iii) $m=-\frac{1}{2}$: general term $=\dfrac{(-\frac{1}{2})(-\frac{3}{2})(-\frac{5}{2})\ldots(-n+\frac{1}{2})}{n!}x^n$

$$=(-1)^n\frac{1\,.\,3\,.\,5\ldots(2n-1)}{2\,.\,4\,.\,6\ldots 2n}x^n$$

and the expansion is:

$$\frac{1}{\sqrt{1+x}}=1-\tfrac{1}{2}x+\frac{1\,.\,3}{2\,.\,4}x^2-\frac{1\,.\,3\,.\,5}{2\,.\,4\,.\,6}x^3+\frac{1\,.\,3\,.\,5\,.\,7}{2\,.\,4\,.\,6\,.\,8}x^4-\ldots$$

$$=1-\tfrac{1}{2}x+\tfrac{3}{8}x^2-\tfrac{5}{16}x^3+\tfrac{35}{128}x^4-\ldots \quad (-1<x<1)$$

A Binomial Series provides a rational approximation to the power $(1+x)^m$ for a given x $(-1<x<1)$. The last case, for example, gives for $x=0{\cdot}1$:

$$\frac{1}{\sqrt{1\cdot 1}}=1-\tfrac{1}{2}(\tfrac{1}{10})+\tfrac{3}{8}(\tfrac{1}{10})^2-\tfrac{5}{16}(\tfrac{1}{10})^3+\tfrac{35}{128}(\tfrac{1}{10})^4-\ldots$$

$$=1-0{\cdot}05+0{\cdot}00375-0{\cdot}00031+0{\cdot}00003-\ldots$$

$$=0{\cdot}9535 \quad \text{to four decimal places.}$$

11.8. Properties of expansions. Given $f(x)$ with derivatives of all orders, Taylor's Series provides the expansion of $f(x)$ as a power series of the form $\Sigma f^{(n)}(0)\dfrac{x^n}{n!}$, for $|x|<r$, where the radius of convergence r has to be found by consideration of the limit of the remainder R_n as $n\to\infty$. The converse situation is: given a power series Σa_nx^n and knowing its radius of convergence r, what is the sum $f(x)$? If $f(x)$ can be found, then $f(x)=\Sigma a_nx^n$ for $|x|<r$. The coefficients a_n are identified: $a_n=\dfrac{f^{(n)}(0)}{n!}$. In practice, it is by no means always possible, given the convergent power series Σa_nx^n, to find its sum $f(x)$ in

terms of known functions. We know $f(x)$ exists, for the sum of a convergent power series must be a function of x. If we cannot find it, we may suspect that we have a new function, something outside the established range of functions. This is, in fact, the situation for the power series of the examples of 11.6. Their sums are not *algebraic* functions; they are new functions to be explored in Chapter 12.

One fact, implicitly assumed above, does serve to simplify matters: if $f(x)=\Sigma a_n x^n$ can be written for certain coefficients $a_0, a_1, a_2, \ldots,$ then this expansion of $f(x)$ is unique. Suppose $f(x)=\Sigma a_n x^n=\Sigma b_n x^n$. Then:

$$f(x)=a_0+a_1x+a_2x^2+\ldots+a_nx^n+x^n(a_{n+1}x+a_{n+2}x^2+\ldots)$$

i.e. $f(x)=a_0+a_1x+a_2x^2+\ldots+a_nx^n+\epsilon(x)x^n$ where $\epsilon(x)\to 0$ as $x\to 0$.

Similarly:

$f(x)=b_0+b_1x+b_2x^2+\ldots+b_nx^n+\eta(x)x^n$ where $\eta(x)\to 0$ as $x\to 0$.

These relations are true for any integral n. Hence:

$$(a_0-b_0)+(a_1-b_1)x+(a_2-b_2)x^2+\ldots+(a_n-b_n)x^n+\{\epsilon(x)-\eta(x)\}x^n=0$$

Let $x\to 0$: $a_0-b_0=0$ i.e. $a_0=b_0$.

Then:

$$(a_1-b_1)+(a_2-b_2)x+\ldots+(a_n-b_n)x^{n-1}+\{\epsilon(x)-\eta(x)\}x^{n-1}=0.$$

Let $x\to 0$: $a_1-b_1=0$ i.e. $a_1=b_1$

and so on. Generally $a_n=b_n$. Q.E.D.

Within its radius of consequence, a power series $\Sigma a_n x^n$ is absolutely convergent. It is a property of absolutely convergent series, Σu_n and Σv_n, that they can be multiplied in just the same way as polynomials:

$$(u_1+u_2+u_3+\ldots)(v_1+v_2+v_3+\ldots)$$
$$=u_1v_1+(u_1v_2+u_2v_1)+(u_1v_3+u_2v_2+u_3v_1)+\ldots.$$

The proof of this result, which is tricky if not difficult, is given in 15.6. It follows that two power series, $\Sigma a_n x^n$ and $\Sigma b_n y^n$, can be so multiplied for all x within both radii of convergence. An example illustrates:

(i) $f(x)=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\ldots+\frac{x^n}{n!}+\ldots$ absolutely convergent, all x.

So: $f(y)=1+y+\frac{y^2}{2!}+\frac{y^3}{3!}+\ldots+\frac{y^n}{n!}+\ldots$ absolutely convergent, all y.

Hence

$$f(x)\times f(y)=\left(1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\ldots\right)\left(1+y+\frac{y^2}{2!}+\frac{y^3}{3!}+\ldots\right)$$

$$=1+(x+y)+\frac{x^2+2xy+y^2}{2!}+\frac{x^3+3x^2y+3xy^2+y^3}{3!}+\ldots$$

$$=1+(x+y)+\frac{(x+y)^2}{2!}+\frac{(x+y)^3}{3!}+\ldots$$

i.e. $f(x)\times f(y)=f(x+y)$.

As a last problem to explore, suppose $f(x)=\Sigma a_n x^n$, absolutely convergent within the radius of convergence r, and ask the questions: can we write $\int f(x)\,dx$ as the sum of the series obtained by integrating $\Sigma a_n x^n$ term by term, and can we write $f'(x)$ as the sum of the series obtained by writing derivatives of $\Sigma a_n x^n$ term by term? In short, can we write the following?

$$\int f(x)\,dx=a_0x+a_1\frac{x^2}{2}+a_2\frac{x^3}{3}+\ldots+a_n\frac{x^{n+1}}{n+1}+\ldots\quad(+\text{constant})$$

$$f'(x)=a_1+2a_2x+3a_3x^2+\ldots+na_nx^{n-1}+\ldots.$$

This is a much more troublesome problem than might appear at first sight. It can only be approached by writing $f(x)=\sum_{s=0}^{n}a_sx^s+R_n$, where R_n is the remainder in Taylor's Series and such that $R_n\to 0$ as $n\to\infty$ for $|x|<r$. Then $\int f(x)\;\;dx=\sum_{s=0}^{n}a_s\frac{x^{s+1}}{s+1}+\int R_n\,dx$. The desired result follows, provided that $\int R_n\,dx\to 0$ as $n\to\infty$. This would seem to be in order, since integrals are 'well-behaved', provided only that the functions concerned are continuous. It can, indeed, be established as correct. On the other hand, in writing $f'(x)=\sum_{s=1}^{n}sa_sx^{s-1}+R_n'(x)$, where $R_n'(x)$ is the derivative of R_n as a function of x, we must expect trouble. We can never be sure of being able to write derivatives even for continuous functions. If we can, we must not expect that $R_n'(x)\to 0$ simply because $R_n\to 0$ as $n\to\infty$. We give this problem up; derivation term by term of a power series is not a safe process.

The first problem, that of integration term by term, is safe enough;

but the result is not easy to prove. We may argue: for any x within the interval $[-r, r]$, $R_n \to 0$ as $n \to \infty$ so that, given ϵ, we can pick N sufficiently large to ensure that $|R_n| \leqslant \epsilon$ for $n > N$. By the Mean Value Theorem for integrals (11.1):

$$\int_a^b R_n\,dx = \Big[R_n\Big]_{x=\alpha}(b-a) \leqslant \epsilon(b-a) \quad \text{for } n > N$$

for any a and b of $[-r, r]$ and for some α $(a < \alpha < b)$. Hence, $\int_a^b R_n\,dx$ can be made as small as we please, by choice of ϵ and N, i.e. $\int_a^b R_n\,dx \to 0$ as $n \to \infty$ for any a and b of $[-r, r]$. Take $b = x$, so that $\int_a^b R_n\,dx = \int R_n\,dx \to 0$ as $n \to \infty$ for x within $[-r, r]$. For such x, $R_n \to 0$ and $\int R_n\,dx \to 0$ as $n \to \infty$. Hence:

$$f(x) = \sum a_n x^n \quad \text{and} \quad \int f(x)\,dx = \sum a_n \frac{x^{n+1}}{n+1} + \text{constant}$$

and integration term by term is a valid process.

There is a difficulty here and it is a subtle one. The remainder R_n depends on x so that, in picking N for $|R_n| \leqslant \epsilon \, (n > N)$ we can only specify N in terms of the particular x we have in mind. Hence N depends on x and this throws out the subsequent line of argument as given. The difficulty disappears if N is independent of x, so that $|R_n| \leqslant \epsilon$ for $n > N$ *and* for all x within $[-r, r]$. In this case, the power series is said to be *uniformly convergent*, i.e. absolutely convergent at all x of $[-r, r]$ in such a way that the choice of ϵ and N do not involve x. The result we need is that, within the radius of convergence, a power series is uniformly as well as absolutely convergent. The result is correct and the proof, which needs some care, is given in 15.6. Accepting it, we can proceed to integrate power series term by term. An example illustrates:

(ii) $$\frac{1}{1+x} = 1 - x + x^2 - x^3 + \dots + (-1)^{n-1}x^{n-1} + \dots \quad (-1 < x < 1).$$

So:

$$\int \frac{dx}{1+x} + \text{constant} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots + (-1)^{n-1}\frac{x^n}{n} + \dots (-1 < x < 1).$$

Take the integral over $[0, x]$ so that the integral and series both vanish when $x=0$ and the constant is zero:

$$\int_0^x \frac{dt}{1+t} = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots + (-1)^{n-1}\frac{x^n}{n} + \ldots \quad (-1<x<1).$$

We start with a geometric series (a case of the Binomial Series of 11.7); we obtain a series examined in example (iii)(a) of 11.6. Both are absolutely convergent for $-1<x<1$. The sum of the first series is $\dfrac{1}{1+x}$. The sum of the second series is not known; if we write it $f(x)$, then $f(x)=\displaystyle\int_0^x \frac{dt}{1+t}$ and we do have a little extra information. In fact, though $\dfrac{1}{1+x}$ is algebraic, $\displaystyle\int_0^x \frac{dt}{1+t}$ is not an algebraic function. It is one of the new functions to be introduced.

11.9. Exercises

1. If $f(x)=x^3-3x^2+2x+1$, show that $f(x)=1$ at $x=0, 1, 2$ and only at these values. Use Rolle's Theorem to show that $f'(\alpha)=0$ for some α $(0<\alpha<1)$ and that $f'(\beta)=0$ for some $\beta\,(1<\beta<2)$. Write $f'(x)$ and deduce that $\alpha=1-\frac{1}{3}\surd 3$ and that $\beta=1+\frac{1}{3}\surd 3$. Show also that $f'(x)=0$ only at these two values.

2. For $f(x)$ of Ex. 1, use the Mean Value theorem to show that $f'(\alpha)=3$ for some $\alpha\,(1<\alpha<3)$, and use $f'(x)$ to show that $\alpha=1+\frac{2}{3}\surd 3$. Is there any other value of x such that $f'(x)=3$?

3. Use (1) and (2) of 11.1 to show that $\sqrt{1+x}=1+\frac{1}{2}x-\frac{1}{8}x^2$ approximately (x small), the remainder being $R=\dfrac{x^3}{16\surd(1+\theta x)^5}$ for some real θ $(0<\theta<1)$. In Taylor's Theorem, take $f(x)=\surd x$ and show that

$$\sqrt{a+h}=\surd a+\frac{h}{2\surd a}-\frac{h^2}{8\surd a^3}+\frac{h^3}{16\surd(a+\theta h)^5}.$$

Hence check the first result.

4. *Taylor's Theorem: alternative form of remainder.* Write $F(x)$ as in the proof of Taylor's Theorem (11.1), with $F'(x)=-\dfrac{(b-x)^n}{n!}f^{(n+1)}(x)$. Show that $\displaystyle\int_a^b \frac{(b-x)^n}{n!}f^{(n+1)}(x)\,dx=-\int_a^b F'(x)\,dx=F(a)$. Put $b=a+h$ and write $x=a+th$ $(0\leqslant t\leqslant 1)$ to deduce that

$$f(a+h)=f(a)+hf'(a)+\frac{h^2}{2!}f''(a)+\ldots+\frac{h^n}{n!}f^{(n)}(a)+R_n$$

where

$$R_n=\frac{h^{n+1}}{n!}\int_0^1 (1-t)^n f^{(n+1)}(a+th)\,dt.$$

Use the Mean Value Theorem for integrals (11.1) to show that

$$R_n = \frac{h^{n+1}}{n!}(1-\theta)^n f^{(n+1)}(a+\theta h) \quad (0<\theta<1)$$

and obtain the remainder form (3) of 11.1 in the case $a=0$, $h=x$.

5. Apply the Mean Value Theorem for integrals to show that $\int_0^1 x^2\,dx = \alpha^2$ for some $\alpha (0 \leqslant \alpha \leqslant 1)$. Evaluate $\int_0^1 x^2\,dx$ and show that $\alpha = \frac{1}{\surd 3}$.

6. *The sign of the derivative.* If $f'(x)>0$ over the interval $[a, b]$, show that $f(x)$ increases continuously over the interval. If $c=f(a)$ and $d=f(b)$, show also that $f^{-1}(x)$ increases continuously over $[c, d]$ with range $[a, b]$.

7. Draw a graph of $y=x^3-3x^2+5$ and locate its single maximum and its single minimum point. Similarly, from a graph of $y=x^3-3x^2+3x+5$, show that this function has only one stationary value, a point of inflexion.

8. Show that $y=5x^3-3x^5$ has a single maximum value (at $x=1$), a single minimum value (at $x=-1$), and in between a third stationary value which is a point of inflexion (at $x=0$).

9. *Maximum profits.* If the revenue $R(x)$ of a firm and its cost of production $C(x)$ both depend on its output x, show that profits are a maximum at output α if $R'(\alpha)=C'(\alpha)$. Interpret as: marginal revenue = marginal cost. Write a sufficient condition for maximum profits. Re-work 10.9 Ex. 16 on this basis.

10. The maximum and minimum values of a variable $z=x^2+y^2$ are sought, where x and y take all real values subject to $4x+3y=5$. Eliminate y, obtaining z as a function of x, and show that $z=1$ is the minimum value and that z has no other extreme value.

*11. *Functions of two variables.* The real variable z is a function of u, where u is the number pair (x, y), x and y real variables. Interpret: given x and y, each from the domain of all real numbers, then z has a unique value $f(x, y)$. Show that the function is represented graphically by a surface in three dimensions, referred to axes $Oxyz$, and interpret as a mapping of points in the plane Oxy onto points on the surface, or onto points on the line Oz. Examine the shape of the surface $z=x^2+y^2$, for all real x and y, showing that cross sections perpendicular to Oz are all circles.

*12. *Partial derivatives.* For $z=f(x, y)$, if $\lim\limits_{h\to 0} \frac{f(x+h, y)-f(x, y)}{h}$ exists, call its value the partial derivative of z with respect to x and write it as $z_x' = f_x'(x, y)$, or as $\frac{\partial z}{\partial x} = \frac{\partial}{\partial x} f(x, y)$. Define the other partial derivative (with respect to y). Interpret $\frac{\partial z}{\partial x}>0$ at (x, y), and similarly $\frac{\partial z}{\partial y}>0$. Show that necessary conditions that z has a local maximum or minimum, for variation of both variables around (x, y), are that $\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y} = 0$. Illustrate with $z=x^2+y^2$, showing that there is only one extreme value (a minimum at $x=0$, $y=0$).

*13. *Constrained maximum and minimum values.* Extreme values of

$z=f(x, y)$ are sought subject to the constraint, the side relation $\phi(x, y)=0$. One method of treatment is that of Ex. 10. Develop the following alternative, called the method of the *Lagrange multiplier* after Lagrange (1736–1813). Write $w=f(x, y)-\lambda\phi(x, y)$ for any multiple λ. The unconstrained extreme value of w corresponds to the constrained extreme value of z. Write $\frac{\partial w}{\partial x}=\frac{\partial w}{\partial y}=0$ and show that a necessary condition for the constrained extreme value sought is the equality of the ratios $f_x'(x, y) : \phi_x'(x, y)$ and $f_y'(x, y) : \phi_y'(x, y)$, each being λ, to be taken in conjunction with $\phi(x, y)=0$.

*14. Apply the necessary condition of Ex. 13 to the problem of Ex. 10, obtaining the same constrained minimum of z. Interpret graphically the maximum (or minimum) of $z=f(x, y)$ subject to a *linear* restraint as the local highest (or lowest) point on a section of the surface $z=f(x, y)$ by a vertical plane (parallel to Oz).

*15. Show that the problem of determining the rectangle of greatest area which can be cut from a circular piece of cardboard of unit radius is equivalent to finding the maximum of $z=4xy$ relative to $x^2+y^2-1=0$ $(x>0, y>0)$. Show that the (maximum) rectangle is a square of side $\sqrt{2}$.

*16. Show that the same x and y give an extreme value of $u=f(x, y)$ relative to $\phi(x, y)=$constant and an extreme value of $v=\phi(x, y)$ relative to $f(x, y)=$ constant. Why would you expect a maximum of u to correspond to a minimum of v? Express the problem of Ex. 15 in alternative ways.

*17. *Joint production.* A firm produces outputs x and y of two commodities, and its given resources limit x and y according to a specific relation $\phi(x, y)=0$ $(x>0, y>0)$. The firm can sell its outputs at given prices, p and q respectively. Show that a necessary condition for maximum revenue is: $\frac{\phi_x'(x, y)}{p}=\frac{\phi_y'(x, y)}{p}$ and interpret in marginal terms. If $\phi(x, y)=x^2+y^2-1$, and if $p : q=4 : 3$, show that maximum revenue is obtained by the outputs $(\frac{4}{5}, \frac{3}{5})$ given by the point P of Fig. 11.9, where one of the lines $4x+3y=$constant touches the circle centre O and of unit radius. Relate this solution to that of Ex. 10, taking note of the result of Ex. 16.

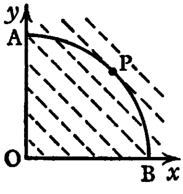


FIG. 11.9

18. *Infinite integrals.* Show that a double limiting process is involved in writing $\int_a^\infty f(x)\,dx$, i.e. first the limit as $n\to\infty$ of the sum of the areas of n rectangles (similar to the sum of an infinite series) and then the limit as the rectangles are refined (more and thinner rectangles).

19. *Integral Test for convergence:* if $f(x)$ is positive, continuous and decreasing for all $x\geqslant 1$, and if $\int_1^\infty f(x)\,dx=L$ is a convergent infinite integral, then Σu_n where $u_n=f(n)$ is convergent to sum $S\leqslant u_1+L$. To prove, write

$v_n = u_n - \int_n^{n+1} f(x)\,dx$ and express $v_n = \int_n^{n+1} \{f(n) - f(x)\}\,dx = f(n) - f(\alpha)$ for $n \leqslant \alpha \leqslant n+1$ by the Mean Value Theorem for integrals. Hence show that $0 \leqslant v_n \leqslant f(n) - f(n+1)$ and that Σv_n is convergent with sum $\leqslant u_1$. Finally $\Sigma v_n = \Sigma u_n - \int_1^\infty f(x)\,dx$ and Σu_n is convergent with sum $\leqslant u_1 + L$.

20. By the comparison theorem (11.4) with $v_n = 1/n$, show that $\sum(1/n^p)$ is not convergent if $0 < p \leqslant 1$, and by the integral test (Ex. 19) show that $\sum(1/n^p)$ is convergent if $p > 1$. Deduce that $1 + \dfrac{1}{2\sqrt{2}} + \dfrac{1}{3\sqrt{3}} + \dots$ is convergent, and that $1 - \dfrac{1}{\sqrt{2}} + \dfrac{1}{\sqrt{3}} - \dfrac{1}{\sqrt{4}} + \dots$ is conditionally convergent.

21. From d'Alembert's test in limit form, establish that $\sum n^k r^n$, for positive k, is convergent $0 < r < 1$ and not convergent $r \geqslant 1$. Deduce that

$$1 + \sqrt{2}\left(\tfrac{3}{4}\right) + \sqrt{3}\left(\tfrac{3}{4}\right)^2 + \dots$$

is convergent.

22. *Speed of convergence.* Given that the common logarithm $\log_{10} 2$ is the sum of the series $0{\cdot}4343(1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots)$, it would appear that more than 200 terms are needed to give $\log_{10} 2$ to two decimal places (0·30). Increase the speed of convergence by writing the terms in pairs: $1 - \frac{1}{2} = \frac{1}{2}$, $\frac{1}{3} - \frac{1}{4} = \frac{1}{12}, \dots$. Show that the nth pair is $1/[2n(2n-1)]$ and that about 20 pairs are needed to get $\log_{10} 2 = 0{\cdot}30$. Examine $\pi = 4(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots)$ similarly.

23. Alternative series for $\log_{10} 2$ are 0·4343 times either:

$$\tfrac{1}{2}\{1 + \tfrac{1}{2}(\tfrac{1}{2}) + \tfrac{1}{3}(\tfrac{1}{2})^2 + \tfrac{1}{4}(\tfrac{1}{2})^3 + \dots\} \quad \text{or} \quad \tfrac{2}{3}\{1 + \tfrac{1}{3}(\tfrac{1}{3})^2 + \tfrac{1}{5}(\tfrac{1}{3})^4 + \tfrac{1}{7}(\tfrac{1}{3})^6 + \dots\}.$$

Show that these converge so rapidly that $\log_{10} 2 = 0{\cdot}30$ (to two decimal places) is got by taking 5 or 6 terms of the first and only two terms of the second series. How many terms for the third decimal place ($\log_{10} 2 = 0{\cdot}301$)?

24. *Recurring decimals.* A recurring decimal is an infinite geometric series, the sum being the corresponding fraction. Illustrate with:

$$0{\cdot}58\dot{3} = \frac{58}{100} + \frac{3}{1000}\left(1 + \frac{1}{10} + \frac{1}{10^2} + \dots\right) = \frac{7}{12}$$

and

$$1{\cdot}\dot{6}\dot{3} = 1 + \frac{63}{100}\left(1 + \frac{1}{10^2} + \frac{1}{10^4} + \dots\right) = \frac{18}{11}.$$

25. Write

$$0{\cdot}\dot{9} = \frac{9}{10}\left(1 + \frac{1}{10} + \frac{1}{10^2} + \dots\right) = 1$$

and deduce that a terminating decimal can be shown as a recurring decimal, e.g. $0{\cdot}64 = 0{\cdot}63\dot{9}$. See also 2.9 Ex. 1.

*26. *Real numbers and decimals.* Illustrate and prove the following series of properties relating non-terminating decimals $0{\cdot}b_1 b_2 \dots b_r \dots$ to real numbers a, $0 < a \leqslant 1$: (i) any such decimal corresponds to one real number and no two decimals correspond to the same real number; (ii) any recurring decimal corresponds to a rational number and conversely; (iii) any non-recurring decimal corresponds to an irrational number and conversely.

*27. *The irrational e.* Define $e=1+1+\frac{1}{2!}+\frac{1}{3!}+\dots$ and suppose e is rational, i.e. $e=\frac{p}{q}$ where p and q are positive integers. If S_q is sum of $(q+1)$ terms of the series, show that

$$\left(\frac{p}{q}-S_q\right)\left(q!\right)=\frac{1}{q+1}+\frac{1}{(q+1)(q+2)}+\dots<\frac{1}{q+1}+\left(\frac{1}{q+1}\right)^2+\dots=\frac{1}{q}.$$

Further, show that $\left(\frac{p}{q}-S_q\right)\left(q!\right)=$positive integer $\left(\not<\frac{1}{q}\right)$. Deduce that e is irrational.

28. Show that $x+\frac{1}{3}x^3+\frac{1}{5}x^5+\dots$, absolutely convergent for $|x|<1$, is not even conditionally convergent when $x=\pm 1$.

*29. *Hypergeometric series:*

$$1+\frac{\alpha}{1}\frac{\beta}{\gamma}x+\frac{\alpha(\alpha+1)}{1\,.\,2}\frac{\beta(\beta+1)}{\gamma(\gamma+1)}x^2+\frac{\alpha(\alpha+1)(\alpha+2)}{1\,.\,2\,.\,3}\frac{\beta(\beta+1)(\beta+2)}{\gamma(\gamma+1)(\gamma+2)}x^3+\dots.$$

Write the term in x^n. If α, β and γ are any constants, other than negative integers, show that the series is a power series absolutely convergent for $|x|<1$. What happens to the series if α, β or γ is a negative integer?

30. If $0<x<1$, write $\frac{1}{\sqrt{1-x^2}}$ as a series in ascending even powers of x. Hence, show that $\frac{1}{\sqrt{0{\cdot}99}}=1{\cdot}00504$ to five decimal places, checking that only three terms are needed.

31. In the product $f(x)\times f(y)$ of example (i) of **11.8**, show that the $(n+1)$th term is:

$$\frac{1}{n!}\left\{x^n+nx^{n-1}y+\frac{n(n-1)}{2!}x^{n-2}y^2+\dots+nxy^{n-1}+y^n\right\}$$

and identify as $\frac{1}{n!}(x+y)^n$ by the Binomial Theorem.

32. Integrate the expansions of $\frac{1}{1+x^2}$ and of $\frac{1}{1-x^2}$ term by term to get

$$\left.\begin{aligned}\int_0^x\frac{dt}{1+t^2}&=x-\frac{x^3}{3}+\frac{x^5}{5}-\frac{x^7}{7}+\dots\\ \int_0^x\frac{dt}{1-t^2}&=x+\frac{x^3}{3}+\frac{x^5}{5}+\frac{x^7}{7}+\dots\end{aligned}\right\}(-1<x<1).$$

33. If $f(x)=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\dots$, show that $\int f(x)\,dx=f(x)+\text{constant}$, i.e. $f(x)$ is such that it is its own derivative. (It is the exponential function of **12.2** below.)

CHAPTER 12

ELEMENTARY FUNCTIONS

12.1. Defining new functions. The specific functions so far considered are all of algebraic form, obtained from a real variable x and appropriate constants by the algebraic processes of addition, subtraction, multiplication, division and root extraction. They include polynomials and ratios of polynomials, the basic element being the power x^n for integral n. They include various expressions involving powers such as $x^r = \sqrt[q]{x^p}$ where r is a rational number p/q $(q>0)$. It is now time to extend the variety of specific functions.

As an illustration, consider example (ii) of 11.8. The function $y=1/1+x$ is of simple algebraic type, continuous and decreasing over the domain $x>-1$. The graph of Fig. 12.1 shows a curve falling smoothly from left to right. The function has an integral: $f(x)=\int_0^x \frac{dt}{1+t}$ where the lower end of the interval of integration is taken so that $f(0)=0$. For given x, $f(x)$ is represented by the shaded area of Fig. 12.1, the area under the curve $y=1/(1+x)$ from 0 to x. In the definition of an integral, we carry out an algebraic process: summing areas of rectangles and proceeding to a limit. We might well expect that the result, the integral $f(x)$, is of algebraic form.

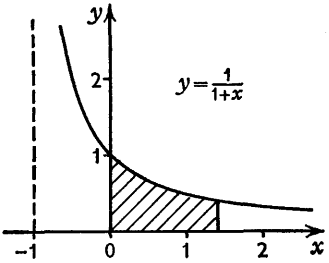


FIG. 12.1

Looking at the problem from another angle, we write the function $1/1+x$ as the sum of the power series $1-x+x^2-x^3+\ldots$ for $|x|<1$ and we get:

$$f(x)=\int_0^x \frac{dt}{1+t}=x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\ldots \quad \text{with } f(0)=0.$$

If x is small, say small enough for x^4 to be neglected in an approximation, then $1/1+x=1-x+x^2-x^3$ and $f(x)=x-\frac{1}{2}x^2+\frac{1}{3}x^3$ approximately, as cubic polynomials. The exact representation of each is obtained by summing n terms of the series and by proceeding to the limit. Again, since we start from an algebraic expression $1/(1+x)$, we might well expect an algebraic result to emerge for $f(x)$.

The surprising fact we now have to face is that our expectation is not justified. The limiting process, to get an integral or the sum of an infinite series, takes us outside the range of algebraic expressions. It produces a function, to be written $\log(1+x)$, of entirely new and non-algebraic type.

There are two ways of defining new functions. One is as the indefinite integral of a known function of x. If the result is not recognised as a function of known type, then it can be written as defining a new function. The other way is as the sum of an infinite power series in ascending powers of x. This sum is again a function of x and, if it is apparently not of known type, then it can be defined as a new function. The two methods usually link up, i.e. a function defined as an integral can be expanded as a power series, and conversely. Clearly we have a choice: we can use one method as the *definition* of a new function, and the other method provides a *property* of the function. For example, we may write $\log(1+x)=\int_0^x \frac{dt}{1+t}$ as a definition and get $\log(1+x)=x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\dots$ as a property; or we may proceed conversely. In either case, we may make a mistake in failing to recognise the 'new' function as an 'old' one. As far as we know at the outset, what we have written as $\log(1+x)$ may turn out to be such an algebraic expression as $\surd(1+x)-1/\surd(1+x)-x^3/24$. Here it is not, but it is always a possibility.* But there's no harm done. If we start by writing $\log(1+x)$ and then find it is really a known function, we can just scrub out the $\log(1+x)$ notation, or indeed we may opt to carry it as a convenient short-hand for the known form.

The choice of which method to adopt as a definition is not so much a question of logic, for each method can be made equally strict. It is more a matter of getting a neat and economical definition from which all desired properties flow easily. The balance lies in favour of using

* The algebraic expression given and $\log(1+x)$ have the same power series up to and including the term in x^3 but not beyond.

an integral as a definition. For, nothing more is needed as a basis than the Fundamental Theorem of the Calculus; if the original function is continuous, the integral exists, is continuous and has a known derivative (the original function). All the new functions of the present chapter are found to stem from just two integrals: $\int_1^x \frac{dt}{t}$ and $\int_0^x \frac{dt}{1+t^2}$. The first gives the exponential, logarithmic, power and hyperbolic functions; the second leads to the circular or trigonometric functions and their inverses.* An outline of this development is given in 15.7 and 15.8.

Against this, the definition of functions as power series requires a more elaborate foundation: the theory of convergence of power series. It is much less neat and economical. In the present development, however, we have already invested a good deal in the construction of the necessary theory (Chapter 11). Further, we find that the expansions of the new functions are of central importance in their manipulation. Consequently, at the cost of a certain mathematical elegance, we follow the method of defining functions as power series.

For the most part we find we can get by with the series:

$$1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\dots$$

together with the similar series:

$$1-\frac{x^2}{2!}+\frac{x^4}{4!}-\frac{x^6}{6!}+\dots \quad \text{and} \quad x-\frac{x^3}{3!}+\frac{x^5}{5!}-\frac{x^7}{7!}+\dots .$$

All are absolutely convergent for all x. A basic constant, due to Euler (1707–83), is a particular case ($x=1$) of the first series. It is denoted e:

$$e=1+1+\frac{1}{2!}+\frac{1}{3!}+\dots=2{\cdot}71828\dots .$$

We shall also make use of the series:

$$x-\frac{x^3}{3}+\frac{x^5}{5}-\frac{x^7}{7}+\dots .$$

* Other new functions can be defined by integrals, e.g. the B and Γ functions of 12.9 Ex. 28 and 29. More generally, new functions can be defined by differential equations, e.g. Bessel functions of 14.9 Ex. 5 below. Many of the new functions are cases of a wide class of functions $F(x; \alpha, \beta, \gamma)$ in three parameters, derived from the

This is absolutely convergent for $-1<x<1$ and conditionally convergent for $x=\pm 1$ (11.6 above). A constant given as a particular case ($x=1$) is:

$$\pi=4(1-\tfrac{1}{3}+\tfrac{1}{5}-\tfrac{1}{7}+\ldots)=3{\cdot}14159\ldots.$$

This turns out to be the constant π of Archimedes (*circa* 250 B.C.), familiar in elementary geometry.

One further step can be taken in this process of unification. Once we admit complex numbers as exponents in the power function, and as the variable in a power series, we can eliminate even the distinctions we have just made. Powers of the constant e with rational, real or complex exponents pull together the whole lot. The range of functions discussed here is unified by means of the exponential function e^x, a function of most remarkable flexibility and power. There is even a link between the basic constant $e=2{\cdot}71828\ldots$ and the other constant $\pi=3{\cdot}14159\ldots$. The relation which turns this trick is: $e^{2\pi i}=1$. This development is perhaps the most amazing, the most exciting, of all mathematics.

12.2. The exponential function. The power series $1+x+\dfrac{x^2}{2!}+\dfrac{x^3}{3!}+\ldots$ is absolutely (and uniformly) convergent for all real x and its sum is a function, the exponential function:*

DEFINITION: *The* **exponential function** *exp x, read 'exponential x', is:* $\exp x=1+x+\dfrac{x^2}{2!}+\dfrac{x^3}{3!}+\ldots$ *defined for all real x.*

The basic property is obtained by multiplication of series:

$$\exp x\times\exp y=\exp(x+y)\quad\text{for all real } x \text{ and } y\ \ldots\ldots\ldots\ldots(1)$$

This is established in example (i) of 11.8, using the result that absolutely convergent series can be multiplied. From the definition: $\exp 0=1$. Hence:

$$\exp x\times\exp(-x)=\exp(x-x)=\exp 0=1$$

hypergeometric series (12.9 Ex. 31). The range of non-algebraic functions used in applied mathematics (e.g. by the physicist or the engineer) is quite extensive; a detailed analysis of them is given by E. T. Whittaker and G. N. Watson: *Modern Analysis* (Cambridge University Press, 1902).

* 'Exponential' is an adjective formed from 'exponent', which derives from the Latin: *ex* = out of, and *pono, ponere* = to place.

$$\left.\begin{aligned}\text{So:}\qquad & \exp(-x)=\frac{1}{\exp x}\\ \text{and:}\qquad & \frac{\exp x}{\exp y}=\exp x\times\exp(-y)=\exp(x-y)\end{aligned}\right\}\quad\ldots\ldots\ldots\ldots(2)$$

Further: $(\exp x)^2=\exp x\times\exp x=\exp 2x$

as another case of the basic result (1). This line can be developed, exactly as in Appendix A. 1, to give any rational power of $\exp x$:

$$(\exp x)^r=\exp(rx)\quad\text{for rational } r\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

The power series for $\exp x$ can be integrated term by term to give the integral of $\exp x$, including an additive and arbitrary constant:

$$\int\exp x\,dx=\text{constant}+\int 1\,dx+\int x\,dx+\int\frac{x^2}{2!}\,dx+\int\frac{x^3}{3!}\,dx+\ldots$$

$$=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\ldots=\exp x$$

where the constant is put equal to 1 and the arbitrary constant dropped for convenience. Inversely: $D(\exp x)=\exp x$. Hence the *standard forms*:

$$D(\exp x)=\exp x\quad\text{and}\quad\int\exp x\,dx=\exp x\quad\ldots\ldots\ldots\ldots(4)$$

These could not be simpler; all derivatives and integrals of $\exp x$ are $\exp x$. They can be extended to $y=\exp u$ where $u=f(x)$. The rules for derivatives of composite functions (10.3) and integration by substitution (10.7) give:

$$D_xy=D_uy\,D_yu\quad\text{and}\quad\int yu'\,dx=\int y\,du$$

i.e. $D\exp f(x)=\exp f(x)f'(x)$ and $\int\exp f(x)f'(x)\,dx=\exp f(x)$...(5)

These two results are seen, directly, to be each the inverse process to the other. As a particular case of (5), take $f(x)=kx$, where k is a constant and $f'(x)=k$:

$$D\exp(kx)=k\exp(kx)\quad\text{and}\quad\int\exp(kx)\,dx=\frac{1}{k}\exp(kx)\ldots\ldots(6)$$

Apart from $\exp 0=1$, the other special value needed is $\exp 1$.

This value is written as the constant e, an irrational number (see 11.9 Ex. 27):

NOTATION: $e=1+1+\frac{1}{2!}+\frac{1}{3!}+\ldots=2{\cdot}71828\ldots$.

This constant turns out to be, not only an irrational, but also a transcendental number. It is not the root of any polynomial equation with rational coefficients. A rational approximation, to any desired number of decimal places, is obtained by adding a sufficient number of terms of the series, as in 11.5 above. The approximation written here is to five decimal places.

From (3), it follows that $(\exp 1)^r=\exp r$ for any rational r, i.e. $e^r=\exp r$, where e^r is a power with a rational exponent in the elementary sense of Appendix A.1. If x is not rational, e^x has no meaning as yet. We are, therefore, free to define it in any way we find convenient. It is now clear what we have to do, i.e. we write $e^x=\exp x$ as a matter of notation:

NOTATION: *the xth power of* $e=e^x=exp\, x=1+x+\frac{x^2}{2!}+\frac{x^3}{3!}+\ldots$ *for any real x.*

If x happens to be rational, then e^x is the ordinary power of elementary algebra, e.g. $e^3=e\times e\times e$ and $e^{\frac{1}{2}}=\sqrt{e}$. If x is not rational, we have simply opted to take e^x as $\exp x$. The reason is seen when the properties (1) and (2) are translated:

$$e^x\times e^y=e^{x+y};\ e^{-x}=1/e^x;\ e^x/e^y=e^{x-y}\ldots\ldots\ldots\ldots\ldots\ldots(7)$$

In other words, e^x satisfies the familiar properties of powers, now for x any real value and not only for x rational.

We now have two alternative notations $\exp x$ and e^x for the same thing. We do not need to keep both of them. We can drop $\exp x$ and stick to e^x, which is clearly more convenient when multiplying and dividing as in (7), as compared with (1) and (2). The standard forms (4) also translate easily:

$$De^x=e^x \quad \text{and} \quad \int e^x\,dx=e^x$$

and similarly for (5) and (6). However, it is convenient to maintain both notations, for the 'exp' form is much easier to write when the exponential of a complicated expression is taken. So $\exp f(x)$ can be used as well as, or instead of, $e^{f(x)}$. For example, in statistics, the normal distribution is:

$$y = y_0 e^{-\frac{1}{2}[(x-a)^2/\sigma^2]} \qquad (a \text{ and } \sigma \text{ parameters})$$

which can be written more easily as $y = y_0 \exp\left\{-\frac{1}{2}\frac{(x-a)^2}{\sigma^2}\right\}$.

12.3. The logarithmic function. The exponential function e^x is continuous, with derivatives and integrals of all orders. If x is rational, then e^x is a power of a positive constant $e > 1$ in the elementary sense of Appendix A.1. Hence, $e^x > 0$ all rational x, and $e^x \to \infty$ as $x \to \infty$, $e^x \to 0$ as $x \to -\infty$, through rational values. By continuity, the same results hold for all real x. The graph of $y = e^x$ is the familiar growth curve shown in Fig. 12.3. The function and curve are increasing for all x, since $De^x = e^x > 0$ all x.

As a continuous and increasing function, $y = e^x$ has an inverse which is also continuous and increasing. Moreover, both the function and its inverse have derivatives and integrals of all orders. The inverse could be described as the 'inverse exponential' function and denoted by 'exp^{-1}'. It is, however, so important in its own right that it is given a separate name: the logarithmic function,* denoted by 'log'. Hence, if $y = e^x$, then $x = \log y$; equally, if $x = e^y$, then $y = \log x$.

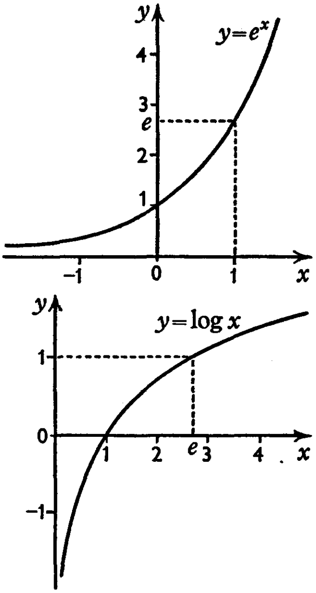


Fig. 12.3

Definition: *The* **logarithmic function** *$y = \log x$, read 'logarithm of x', is the inverse of the exponential function:*

$$\textit{if } x = e^y, \textit{ then } y = \log x \quad \textit{defined for all } x > 0.$$

The graph of $y = \log x$ is simply that of $x = e^y$. Hence, given the graph of $y = e^x$, interchange of axes produces the graph of $y = \log x$, as in Fig. 12.3. Two particular values are to be noted:

$$\log 1 = 0 \quad \text{and} \quad \log e = 1.$$

* 'Logarithm' is derived from the Greek: *logos* = reckoning, and *arithmos* = number.

Hence, as an increasing function, $y = \log x$ is negative for $0 < x < 1$ and positive for $x > 1$. It is not defined for $x \leqslant 0$.

Properties of logarithms are obtained as direct reflections of properties of exponentials. Let $u = \log x$ and $v = \log y$, so that $x = e^u$ and $y = e^v$. Then:

$$xy = e^u \times e^v = e^{u+v} \quad \text{i.e. } \log xy = u + v = \log x + \log y.$$

Similar results for $1/x$ and for x/y are obtained. These results are of the greatest importance, both in theory and practice:

$$\log xy = \log x + \log y; \ \log (1/x) = -\log x; \ \log (x/y) = \log x - \log y \ldots(1)$$

for any positive real values of x and y. Two other results, of a partial kind and subject to later extension (12.4 below), can be usefully set out. From (1), $\log x^2 = \log x + \log x = 2 \log x$; this extends to $\log x^n = n \log x$. Further, if $u = \log (e^x)$, then $e^u = e^x$ and $u = x$, i.e. $\log (e^x) = x$. So:

$$\log x^n = n \log x \ (n \text{ positive integer}) \quad \text{and} \quad \log e^x = x \ldots\ldots\ldots\ldots(2)$$

The derivative of $\log x$ is obtained from that of e^x by the rule for inverse functions (10.3). If $y = \log x$, then $x = e^y$ and:

$$D_x y = 1/D_y x = 1/D_y e^y = 1/e^y = 1/x.$$

Hence the *standard form* and its extension by the composite function rule:

$$D \log x = \frac{1}{x} \text{ for } x > 0 \quad \text{and} \quad D \log f(x) = \frac{f'(x)}{f(x)} \text{ for } f(x) > 0 \ldots\ldots(3)$$

It is not profitable at this stage to seek the integral of $\log x$, though the integral exists. On the other hand, the inverse of (3) is the *standard form*:

$$\int \frac{1}{x}\, dx = \log x \text{ for } x > 0 \quad \text{and} \quad \int \frac{f'(x)}{f(x)}\, dx = \log f(x) \text{ for } f(x) > 0 \ldots\ldots(4)$$

An arbitrary constant is to be added to (4). Alternatively, a lower end of the interval of integration can be specified. Since $\log 1 = 0$, we have:

$$\int_1^x \frac{1}{t}\, dt = \log x \quad (x > 0).$$

This form of (4) is the appropriate (alternative) definition of $\log x$ as an integral. If this definition is adopted, the exponential function comes as the inverse of the logarithmic function, and not (as here) the other way round. The same integral gives a new form (and alternative definition) for e: $\int_1^e \frac{dx}{x} = \log e = 1$. It is also to be noticed that (4) completes the standard form already given (10.7):

$$\int x^r\,dx = \frac{x^{r+1}}{r+1} \quad (r \neq -1).$$

The derivative $f'(x)$ measures the rate of change of $f(x)$, from the definition in 10.2. The units of measurement are so much of $f(x)$ per unit of x, e.g. velocity as the rate of change of distance over time can be measured in miles per hour or in feet per second. Consider the related concept of the proportionate rate of change, i.e. the rate of change of $f(x)$ as a proportion or percentage of $f(x)$ itself:

DEFINITION: *The* **proportionate rate of change** *of* $y=f(x)$ *is:*

$$\frac{1}{y}\,Dy = \frac{f'(x)}{f(x)} = \operatorname*{Lim}_{h\to 0} \frac{f(x+h)-f(x)}{hf(x)}.$$

Another notation often used is $\frac{1}{y}\frac{dy}{dx} = \operatorname*{Lim}_{\Delta x\to 0} \frac{\Delta y}{\Delta x}$, where Δx and Δy are corresponding increments in x and y. Standard form (3) gives:

$$\text{Proportionate rate of change } \frac{1}{y}\,Dy = D\log y \quad (y>0).$$

Hence, the rate of change of y is given by the derivative of y and the proportionate rate of change by the derivative of $\log y$.

The units used for measuring y appear in the rate of change but are eliminated in the proportionate rate of change. For example, if $y=2x^2$ where x is time in years, then $Dy=4x=16$ and $\frac{1}{y}Dy=\frac{2}{x}=\frac{1}{2}$ at $x=4$, i.e. y is then increasing by 16 units per year and by 50 per cent per year.

For the exponential function $y=y_0e^{rx}$, where y_0 is positive (the value at $x=0$) and where r is a positive constant:

$$\log y = \log y_0 + rx \quad \text{and} \quad \frac{1}{y}\,Dy = D\log y = r.$$

The function $y=y_0e^{rx}$ has the property that y grows at a constant proportionate rate r, i.e. steadily at $100r$ per cent per unit of x. The converse is also true (12.9 Ex. 9). For this reason, the exponential function has a wide range of applications, in such problems as compound interest and population growth (12.9 Ex. 13 and 15).

It remains to write the expansion of the logarithmic function. This is done, not for $\log x$, but for $\log(1+x)$. The reason is that $\log(1+x)=0$ at $x=0$ but $\log x$ is not defined at $x=0$. Hence $\log(1+x)$ is defined in the interval $-1<x<1$ around $x=0$. To expand $\log(1+x)$, we repeat what we said in 12.1:

$$1/1+x=1-x+x^2-x^3+\dots \qquad -1<x<1$$

and $$\log(1+x)=\int_0^x \frac{dt}{1+t}=x-\frac{x^2}{2}+\frac{x^3}{3}-\frac{x^4}{4}+\dots \qquad -1<x<1 \quad \dots\dots\dots(5)$$

The expansion (5) is absolutely convergent in the interval shown; it is also conditionally convergent at $x=1$, as found in 11.6 above. However, the series (5) is not convergent at $x=-1$, a reflection of the fact that log 0 is not defined.

The logarithmic function can be used to derive an important limit:

THEOREM: $e^x=\lim\limits_{n\to\infty}\left(1+\frac{x}{n}\right)^n$ *for any real x and integral n.*

Proof: in $D\log f(x)=f'(x)/f(x)$, write $f(x)=1+ux$, $f'(x)=u$, where u is any given real value. So:

$$D\log(1+ux)=\frac{u}{1+ux} \quad \text{and} \quad \Big[D\log(1+ux)\Big]_{x=0}=u.$$

By the definition of the derivative (at $x=0$):

$$\frac{\log(1+uh)-\log 1}{h}=\frac{1}{h}\log(1+uh)\to u \quad \text{as } h\to 0.$$

Write $h=\frac{1}{n}$ and let $n\to\infty$ through integral values:

$$\lim_{n\to\infty} n\log\left(1+\frac{u}{n}\right)=u.$$

By (2) above: $\lim\limits_{n\to\infty}\log\left(1+\frac{u}{n}\right)^n=\log e^u$.

Since the logarithmic function is continuous, this implies that:

$$\lim_{n\to\infty}\left(1+\frac{u}{n}\right)^n=e^u$$

which is the result to be proved, on switching from u to x. Q.E.D.

Hence, the exponential function is the limit of an algebraic form:

$$\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n=e^x \quad\text{and}\quad \lim_{n\to\infty}\left(1-\frac{x}{n}\right)^n=e^{-x} \;\ldots\ldots\ldots\ldots(6)$$

As a particular case, (6) gives e as a limit: $\lim\limits_{n\to\infty}\left(1+\frac{1}{n}\right)^n=e$. Again we see how a non-algebraic function can be got by first writing an algebraic expression and then by proceeding to a limit.

12.4. Power functions. The notation e^x for the xth power of e is most useful. It agrees with the ordinary concept of a power when the exponent is *rational* and it satisfies the basic property of powers ($e^x\times e^y=e^{x+y}$) for all *real* exponents. This is the aspect of the exponential function now to be developed.

Consider a^b where a and b are any real numbers. So far as we have defined this expression only in two cases: when b is rational and a a real number; when $a=e$ and b a real number. It can now be extended to all cases, provided only that $a>0$. If a is real and positive, $\log a$ is defined and so is $e^{b\log a}$ for any real b. This expression provides the definition of the power a^b:

DEFINITION: *The* **power** a^b *for any real* $a>0$ *and any real* b *is:*

$$a^b=e^{b\log a}.$$

Since $\log e^x=x$, so $\log a^b=\log(e^{b\log a})=b\log a$. The definition, therefore, is equivalent to the property:

$$\log a^b=b\log a \;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

The two results of 12.3 are included in and so extended by (1):

$$\log a^n=n\log a\ (n\text{ positive integer})\quad\text{and}\quad \log e^x=x.$$

The second follows from $\log e=1$. Further the basic properties of powers extend at once to a^b. For:

$$a^x\times a^y=e^{x\log a}\times e^{y\log a}=e^{(x+y)\log a}=a^{x+y}.$$

Similarly: $(ab)^x=a^xb^x$. Hence for $a>0$ and $b>0$ and real x and y:

$$a^x\times a^y=a^{x+y}\quad\text{and}\quad (ab)^x=a^xb^x \;\ldots\ldots\ldots\ldots\ldots(2)$$

Particular cases of (2) are: $a^x/a^y=a^{x-y}$ and $1/a^x=a^{-x}$.

An extension of the concept of a logarithm is implied by the development of powers of e into powers of $a>0$. If $x=a^y=e^{y\log a}$

then $\log x = y \log a$, i.e. $y = \dfrac{\log x}{\log a}$ is the inverse of $x = a^y$. As a matter of notation, this inverse is written $\log_a x$, read 'logarithm of x to the base a':

NOTATION: *If $x = a^y$, then $y = \log_a x = \dfrac{\log x}{\log a}$ and a is called the* **base** *of the logarithm $\log_a x$.*

As a check, write $a = e$. So, if $x = e^y$, then $y = \log_e x = \dfrac{\log x}{\log e} = \log x$. There is agreement; $\log x$ is simply $\log_e x$ and the basic concept of a logarithm (12.3) is that of a logarithm to base e. We continue to use $\log x$ instead of $\log_e x$, suppressing the base only when it is e.

Hence, for logarithms to any positive base a, the basic property is:

$$\log_a x = \frac{\log x}{\log a} \quad\text{.....................................}(3)$$

which simply shows how any logarithm $\log_a x$ to base a can be written in terms of logarithms to base e. Indeed, (3) shows that the switch from $\log x$ to $\log_a x$ or conversely is no more than a change of unit:

$$\log_a x = \left(\frac{1}{\log a}\right) \log x \quad \text{and} \quad \log x = (\log a) \log_a x.$$

Logarithms to base e are all multiplied by the constant $\left(\dfrac{1}{\log a}\right)$ to get corresponding logarithms to base a, and similarly for the reverse switch. As a result, all the properties of logarithms, (1) of 12.3 and (1) above, carry over:

$$\left.\begin{array}{c} \log_a xy = \log_a x + \log_a y \;;\; \log_a \dfrac{1}{x} = -\log_a x \;;\; \log_a \dfrac{x}{y} = \log_a x - \log_a y \\ \text{and} \;\; \log_a x^b = b \log_a x \end{array}\right\}(4)$$

Logarithms to base e are called *natural* or *Naperian logarithms* after Napier (1550–1617); they are used here unless the contrary is specified. The logarithms used in arithmetical work are to the convenient base 10 and they are called *common logarithms*. These are a re-scaling of natural logarithms:

$$\log_{10} x = \left(\frac{1}{\log 10}\right) \log x = 0{\cdot}43429 \ldots \log x.$$

Tables are available for common and natural logarithms, giving $\log_{10} x$ and $\log x$ for various values of x. It is not the purpose of the present text to give an account of the use of logarithms in practical computations. What has been done is to fit the ordinary concept of logarithms into the strict development of exponential and logarithmic functions. The use of $y = \log x$, where x and y are any real numbers ($x > 0$) and not only rationals, is now justified. Indeed it is only in the present context that we can justify the everyday use of logarithms.

Two power functions are obtained from the definition of a^b, according to which of the pair of real numbers a and b is given and which is the variable. The *power function* $y = a^x$, for a real constant $a > 0$, is defined on the domain of all real x. It is a development of the exponential function:

$$y = a^x = e^{x \log a}$$

with a graph which is the same as that of $y = e^x$ except for a re-scaling of the variable x. For: $a^{x'} = e^x$ if $x' = x \log a$. Properties (2) above are those to have in mind in handling this power function. The inverse function is $y = \log_a x$, with a graph which comes from that of $y = \log x$ by a re-scaling of the variable y. For: if $y' = \log_a x$ and $y = \log x$, then $y' = \left(\frac{1}{\log a}\right) y$ by property (3). Properties (4) are relevant for $y = \log_a x$.

Derivatives of a^x and $\log_a x$ come from the relationship of these functions to the exponential and logarithmic functions:

$$Da^x = D(e^{x \log a}) = (\log a)e^{x \log a} = a^x \log a$$

$$D \log_a x = \frac{1}{\log a} D \log x = \frac{1}{x \log a}.$$

Also: $$\int a^x \, dx = \int e^{x \log a} \, dx = \frac{e^{x \log a}}{\log a} = \frac{a^x}{\log a}.$$

The expansions are to be obtained likewise:

$$a^x = e^{x \log a} = 1 + (\log a)x + \frac{(\log a)^2}{2!} x^2 + \frac{(\log a)^3}{3!} x^3 + \dots \quad (\text{all } x)$$

$$\log_a (1 + x) = \frac{\log (1 + x)}{\log a} = \frac{1}{\log a}\left(x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots\right) \quad (-1 < x < 1).$$

The other *power function* is $y=x^a$ for a real constant a, defined on the domain of real $x>0$. This is also a modification of the exponential:

$$y=x^a=e^{a\log x}$$

i.e. it is an exponential function giving y in terms of $\log x$. Over the domain $x>0$, $y=x^a$ is continuous and increasing; its inverse is $y=x^{1/a}$, another of the same group of functions. For example, $y=x^2$ and $y=\surd x$ are two particular cases, one the inverse of the other ($x>0$). When a is irrational, the function x^a is not algebraic. When a is rational, we arrive back at an algebraic function x^a. Derivatives and integrals follow (12.9 Ex. 17):

$$D(x^a)=ax^{a-1} \quad \text{and} \quad \int x^a\,dx=\frac{x^{a+1}}{a+1} \quad (a\neq -1).$$

So does the expansion of $(1+x)^a$:

$$(1+x)^a=1+ax+\frac{a(a-1)}{2!}x^2+\frac{a(a-1)(a-2)}{3!}x^3+\dots \quad (-1<x<1)$$

an extension of the *Binomial Series* of 11.7 above.

Two very useful limits can be established:

THEOREM: $\dfrac{\log x}{x^a}$ *and* $\dfrac{x^a}{e^x}$ *both tend to zero as* $x\to\infty$, *for a constant* $a>0$.

Proof: Take $x>0$ and n any positive integer, so that

$$e^x=1+x+\frac{x^2}{2!}+\dots+\frac{x^n}{n!}+\dots>\frac{x^n}{n!}$$

and $$0<\frac{x^a}{e^x}<\frac{n!x^a}{x^n}=\frac{n!}{x^{n-a}} \quad (a>0).$$

Choose $n>a$ so that $\dfrac{1}{x^{n-a}}\to 0$ as $x\to\infty$.

Hence: $$\frac{x^a}{e^x}\to 0 \text{ as } x\to\infty \quad (a>0).$$

Now write: $$\frac{y^b}{e^y}\to 0 \text{ as } y\to\infty \quad (b>0).$$

Put $y=\log x$ and $b=1/a$. Then $x=e^y\to\infty$ as $y\to\infty$

and $$\frac{(\log x)^{1/a}}{x}\to 0 \quad \text{as} \quad x\to\infty \quad (a>0)$$

i.e. $$\left\{\frac{(\log x)^{1/a}}{x}\right\}^a\to 0 \quad \text{as} \quad x\to\infty \quad (a>0)$$

i.e. $\quad \dfrac{\log x}{x^a} \to 0 \quad \text{as} \quad x \to \infty \quad (a>0).$ Q.E.D.

Combining the two results, we conclude that $\log x$; x^a; e^x are in ascending order of magnitude (for large x), all increasing indefinitely with x. In other words, x^a tends to infinity faster than $\log x$, and e^x tends to infinity faster than x^a or $\log x$, as $x \to \infty$. Hence, the ratio of a later to an earlier member of the ordered set of three, e.g. $\dfrac{e^x}{x^a} \to \infty$ as $x \to \infty$; the ratio of an earlier to a later member, e.g. $\dfrac{\log x}{x^a} \to 0$ as $x \to \infty$. The graphs of $y = \log x$, $y = x^2$ and $y = e^x$ illustrate this important relationship between the functions (Fig. 12.4). The relative positions are preserved if (e.g.) x^3 or $\sqrt{x}$ is substituted for x^2.

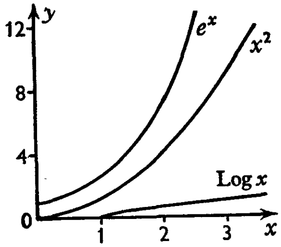


FIG. 12.4

12.5. Circular functions. The power series $1 - \dfrac{x^2}{2!} + \dfrac{x^4}{4!} - \dfrac{x^6}{6!} + \ldots$ and $x - \dfrac{x^3}{3!} + \dfrac{x^5}{5!} - \dfrac{x^7}{7!} + \ldots$ are absolutely (and uniformly) convergent for a real x. As functions of x, the sums of these series define the circular functions.

DEFINITION: *The* **circular functions** *cos x and sin x are defined for all real x:*

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \ldots; \quad \sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$$

where cos x is read 'cosine x' and sin x is read 'sine x'.

These functions are connected one with the other.* Being absolutely convergent for all x, the series can be multiplied to give:

$$\sin x \sin y = \left(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \ldots\right)\left(y - \frac{y^3}{3!} + \frac{y^5}{5!} - \ldots\right)$$

$$= xy - \left(\frac{x^3 y}{3!} + \frac{xy^3}{3!}\right) + \left(\frac{x^3 y^3}{3!3!} + \frac{x^5 y}{5!} + \frac{xy^5}{5!}\right) - \ldots$$

$$= \frac{2xy}{2!} - \frac{4x^3 y + 4xy^3}{4!} + \frac{6x^5 y + 20x^3 y^3 + 6xy^5}{6!} - \ldots$$

* 'Sine' is derived from the Latin: *sinus* = curve. 'Cosine' is the complementary term: cos x equals sin y where x and y are complementary angles (adding to a right angle), see Fig. 12.7c below.

and: $$\cos x \cos y = \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \ldots\right)\left(1 - \frac{y^2}{2!} + \frac{y^4}{4!} - \ldots\right)$$

$$= 1 - \left(\frac{x^2}{2!} + \frac{y^2}{2!}\right) + \left(\frac{x^2y^2}{2!2!} + \frac{x^4}{4!} + \frac{y^4}{4!}\right) - \left(\frac{x^6}{6!} + \frac{x^4y^2}{4!2!} + \frac{x^2y^4}{2!4!} + \frac{y^6}{6!}\right) + \ldots$$

$$= 1 - \frac{x^2 + y^2}{2!} + \frac{x^4 + 6x^2y^2 + y^4}{4!} - \frac{x^6 + 15x^4y^2 + 15x^2y^4 + y^6}{6!} + \ldots$$

Put $x = y$ and write $\sin^2 x$ for $(\sin x)^2$, $\cos^2 x$ for $(\cos x)^2$:

$$\sin^2 x = x^2 - \frac{x^4}{3} + \frac{2}{45}x^6 - \ldots \quad \text{and} \quad \cos^2 x = 1 - x^2 + \frac{x^4}{3} - \frac{2}{45}x^6 + \ldots$$
$$= 1 - \sin^2 x.$$

Hence the first basic property of circular functions:

$$\sin^2 x + \cos^2 x = 1 \quad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

We do not, strictly, need both functions, since one is given in terms of the other: $\sin x = \pm\sqrt{(1 - \cos^2 x)}$, $\cos x = \pm\sqrt{(1 - \sin^2 x)}$. But these are awkward relations and it is convenient to carry both functions, with the relation (1) between them.

Further, subtracting the series for $\sin x \sin y$ and $\cos x \cos y$, we have:

$$\cos x \cos y - \sin x \sin y$$
$$= 1 - \frac{x^2 + 2xy + y^2}{2!} + \frac{x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4}{4!}$$
$$- \frac{x^6 + 6x^5y + 15x^4y^2 + 20x^3y^3 + 15x^2y^4 + 6xy^5 + y^6}{6!} + \ldots$$
$$= 1 - \frac{(x+y)^2}{2!} + \frac{(x+y)^4}{4!} - \frac{(x+y)^6}{6!} + \ldots$$
$$= \cos(x+y).$$

In a similar way (12.9 Ex. 19 and 20):

$$\sin x \cos y + \cos x \sin y = \sin(x+y).$$

Hence, the second basic property of circular functions gives *addition formulae*:

$$\left.\begin{aligned} \cos(x+y) &= \cos x \cos y - \sin x \sin y \\ \sin(x+y) &= \sin x \cos y + \cos x \sin y \end{aligned}\right\} \ldots\ldots\ldots\ldots(2)$$

From the definition, $\cos x = 1$ and $\sin x = 0$ at $x = 0$. Further:

$$\frac{\sin x}{x} = 1 - \frac{x^2}{3!} + \frac{x^4}{5!} - \frac{x^6}{7!} + \ldots \to 1 \text{ as } x \to 0.$$

So: $$\cos 0 = 1;\ \sin 0 = 0;\ \frac{\sin x}{x} \to 1 \text{ as } x \to 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

The power series for $\cos x$ and $\sin x$ can be integrated term by term:

$$\int \cos x\,dx = \text{constant} + \int 1\,dx - \frac{1}{2!}\int x^2\,dx + \frac{1}{4!}\int x^4\,dx - \frac{1}{6!}\int x^6\,dx + \ldots$$

$$= x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots \quad (\text{constant} = 0)$$

$$= \sin x$$

and

$$\int \sin x\,dx = \text{constant} + \int x\,dx - \frac{1}{3!}\int x^3\,dx + \frac{1}{5!}\int x^5\,dx - \frac{1}{7!}\int x^7\,dx + \ldots$$

$$= -1 + \frac{x^2}{2!} - \frac{x^4}{4!} + \frac{x^6}{6!} - \frac{x^8}{8!} + \ldots \quad (\text{constant} = -1)$$

$$= -\cos x$$

where, apart from introducing particular constants, the arbitrary constants are dropped for convenience (as usual with indefinite integrals). By reversing the integration process in $\int \sin x\,dx = -\cos x$, we get the derivative: $D \cos x = -\sin x$. Similarly: $D \sin x = \cos x$. Hence, the *standard forms*:

$$\left.\begin{array}{l} D\cos x = -\sin x;\ D\sin x = \cos x \\ \int \cos x\,dx = \sin x;\ \int \sin x\,dx = -\cos x \end{array}\right\} \ldots\ldots\ldots\ldots\ldots (4)$$

One further derivative is needed:

$$D\left(\frac{\sin x}{\cos x}\right) = \frac{\cos x\,D\sin x - \sin x\,D\cos x}{\cos^2 x} = \frac{\cos^2 x + \sin^2 x}{\cos^2 x}$$

i.e. $$D\left(\frac{\sin x}{\cos x}\right) = 1 + \left(\frac{\sin x}{\cos x}\right)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

As a particular case of the geometric series, and of the Binomial Series of 11.7, the following power series is absolutely convergent for $-1 < x < 1$:

$$1/1 + x^2 = 1 - x^2 + x^4 - x^6 + \ldots$$

On integrating term by term:

$$\int \frac{dx}{1+x^2} = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots$$

which is absolutely convergent for $-1<x<1$ and conditionally convergent also for $x=\pm 1$ (see 11.6). The appropriate range of integration here is from 0 to x, so that the integral becomes zero at $x=0$. Hence, consider $\int_0^x \frac{dt}{1+t^2}$ which is defined for all real x and which is the sum of the power series above for $-1<x<1$. A new function is so defined, called the 'inverse tangent' and denoted '$\tan^{-1}$'. The reason for this odd notation appears later.

Definition: *The* **inverse tangent** *function $tan^{-1} x$ is:*

$$tan^{-1}\, x = \int_0^x \frac{dt}{1+t^2} \quad \textit{defined for all real } x$$

$$= x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots \quad \textit{for } -1 \leqslant x \leqslant 1.$$

From the definition as an integral, the *standard form* follows:

$$D \tan^{-1} x = 1/1 + x^2 \quad \dots\dots\dots\dots\dots\dots(6)$$

Since $\tan^{-1} 0 = 0$ and $D \tan^{-1} x > 0$, all x, $y = \tan^{-1} x$ is a continuous and increasing function of x such that $\tan^{-1} x < 0$ for $x < 0$ and $\tan^{-1} x > 0$ for $x > 0$.

The inverse tangent function, continuous and increasing, has an inverse which is also continuous and increasing and which is called the tangent function,* denoted 'tan'. Hence, if $y = \tan^{-1} x$, then $x = \tan y$; equally, if $x = \tan^{-1} y$, then $y = \tan x$. The domain of the tangent function is the range of the inverse tangent function, to be determined but known to be some interval around $x=0$.

Definition: *The* **tangent function** *$y = tan\, x$ is the inverse of the inverse tangent: if $x = tan^{-1}\, y$, then $y = tan\, x$ for some interval around $x=0$.*

The interval is the range of $\tan^{-1} y$ as y takes all real values.

* The label is the same as for the 'tangent' to a curve, derived from the Latin: *tango, tangere* = to touch. The slope of the tangent to a curve is tan α, where α is the angle the tangent makes with the horizontal.

The derivative of tan x follows from the inverse function rule (10.3). If $y = \tan x$, so that $x = \tan^{-1} y$ and $D_y x = 1/1 + y^2$ by (6), then:

$$D_x y = 1/D_y x = 1 + y^2.$$

Again write $\tan^2 x$ for $(\tan x)^2$. Hence, the *standard form*:

$$D \tan x = 1 + \tan^2 x \quad \text{.........................}(7)$$

The properties of $y = \tan x$ are similar to those of the inverse tangent. Since $\tan 0 = 0$ and $D \tan x > 0$, $y = \tan x$ is continuous and increasing (over its domain) such that $\tan x < 0$ for $x < 0$ and $\tan x > 0$ for $x > 0$.

Now, check the derivative (7) against the derivative (5): tan x, as now defined, turns out to be the ratio of sin x to cos x, as previously defined. In this case, therefore, we have not got a new function at all; it is expressed in terms of functions already known. We do not need a new notation. Having defined sin x and cos x, then $\tan x = \dfrac{\sin x}{\cos x}$ and $\tan^{-1} x$ is the inverse. But, just as we carry both cos x and sin x, so we opt to use tan x subject to:

$$\tan x = \sin x/\cos x \quad \text{............................}(8)$$

There are three circular functions (cos x, sin x, tan x) related by (1) and (8). The inverse, $\tan^{-1} x$, is given by the integral and series written.

In the inverse tangent $y = \tan^{-1} x$, consider the particular value taken by y when $x = 1$, i.e. the sum of the series $1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \dots$. As a matter of notation:

NOTATION: $\frac{1}{4}\pi = \tan^{-1} 1 = \int_0^1 \frac{dx}{1+x^2}$, *so that:*

$$\pi = 4\left(1 - \tfrac{1}{3} + \tfrac{1}{5} - \tfrac{1}{7} + \dots\right) = 3{\cdot}14159\dots.$$

This notation, in effect, defines the constant π. It remains to identify it with the familiar π of elementary geometry (12.7 below). For the moment, we simply have:

$$\tan^{-1} 1 = \pi/4 \quad \text{and} \quad \tan(\pi/4) = 1. \quad \text{.................}(9)$$

The results (3) and (9) summarise the particular values we know so far:

$$\cos x = 1,\ \sin x = 0,\ \tan x = 0 \quad \text{at } x = 0$$
$$\sin x/\cos x = \tan x = 1 \qquad \text{at } x = \pi/4.$$

Further values, and the graphs of the circular functions, are left over until the relationship with the trigonometric ratios is established (12.7).

12.6. Complex exponents. The exponential e^r, r rational, is an ordinary power of the constant e (e.g. $e^3 = e \times e \times e$ and $e^{\frac{1}{2}} = \surd e$). The extension of e^r to e^x, when x is a real variable, is achieved by the device of taking e^x as the sum of a power series. Here e^x has no meaning apart from the power series (except when x happens to be rational) but it does satisfy the essential rule: $e^x \times e^y = e^{x+y}$. It would seem natural to make a further extension from e^x· to e^z, where z is a complex number $x + iy$. The extension, in fact, can be achieved without introducing any really new ideas and it is a very important one. Our aim is to write:

$$Z = e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \dots\dots\dots\dots\dots\dots\dots\dots (1)$$

where $z = x + iy$ is a variable complex number and where $Z = X + iY$ is the complex value obtained from z by taking the exponential e^z. There are two matters to consider in an appreciation of (1).

One point is that $Z = e^z$ is a particular case of a function of a complex variable $Z = F(z)$, as developed as a 'conformal transformation' in 7.6. Such a function involves the double process of 'equating real and imaginary parts'. Specifically, when the operation represented by $F(z)$ is performed on $z = x + iy$, a complex value emerges which depends on x and y:

$$F(z) = \phi(x, y) + i\psi(x, y)$$

where the form of F determines what form the expressions ϕ and ψ take' This is $Z = X + iY$ so that $X = \phi(x, y)$ and $Y = \psi(x, y)$ is the conformal transformation. The function $F(z)$ implies a *pair* of (conformal) relations, one from the equation of 'real' parts, the other from the 'imaginary' parts.

The other point concerns a power series in a complex variable, of the kind written in (1). There is no difficulty here. Suppose Σu_n and Σv_n are two (convergent) series and write $w_n = u_n + iv_n$ as a complex number. Then the sum Σw_n of the series of complex terms w_n simply stands for $\Sigma u_n + i\Sigma v_n$:

NOTATION: *if $w_n = u_n + iv_n$, write $\Sigma w_n = \Sigma u_n + i\Sigma v_n$ and call Σw_n the sum of the series of complex terms w_n.*

Again, in Σw_n, the 'real' and 'imaginary' parts are added separately.

The definition of absolute convergence (11.5) extends: Σw_n is *absolutely convergent* if the series of real positive terms $\Sigma \mid w_n \mid$ is convergent, where $w_n = u_n + iv_n$ and $\mid w_n \mid = \surd(u_n^2 + v_n^2)$. The essential result is:

THEOREM: *$\Sigma w_n = \Sigma(u_n + iv_n)$ is absolutely convergent if and only if Σu_n and Σv_n are both absolutely convergent.*

Proof: If Σw_n is absolutely convergent, then $\Sigma \mid w_n \mid$ is convergent. But $\mid u_n \mid \leqslant \surd(u_n^2 + v_n^2) = \mid w_n \mid$ and similarly for $\mid v_n \mid$. Hence both $\Sigma \mid u_n \mid$ and $\Sigma \mid v_n \mid$ are convergent, i.e. both Σu_n and Σv_n are absolutely convergent. Conversely, if Σu_n and Σv_n are absolutely convergent, then $\Sigma \mid u_n \mid$ and $\Sigma \mid v_n \mid$ are convergent series of positive terms, and so is $\Sigma(\mid u_n \mid + \mid y_n \mid)$. Since $\mid w_n \mid = \surd(u_n^2 + v_n^2) \leqslant \mid u_n \mid + \mid v_n \mid$, $\Sigma \mid w_n \mid$ is convergent and Σw_n is absolutely convergent. Q.E.D.

Consequently, as long as Σu_n and Σv_n are absolutely convergent, we can deal with the absolutely convergent series Σw_n of complex terms $w_n = u_n + iv_n$. In particular, such series can be multiplied since the property of 11.8 continues to hold. All this applies to a power series $\Sigma a_n z^n$ where $z = x + iy$. The general term:

$$a_n z^n = u_n + iv_n \quad \text{where } u_n \text{ and } v_n \text{ depend on } x \text{ and } y.$$

If Σu_n and Σv_n are absolutely convergent (real) series, then $\Sigma a_n z^n$ is absolutely convergent. In this case:

$$Z = \Sigma a_n z^n = \Sigma(u_n + iv_n)$$

means $$Z = X + iY \text{ where } X = \Sigma u_n \text{ and } Y = \Sigma v_n.$$

Absolutely convergent power series can always be multiplied together.

The power series $1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \dots$ is absolutely convergent for all z. To prove: write $w_n = \frac{z^n}{n!}$ so that $\mid w_n \mid = \frac{\mid z \mid^n}{n!}$ where $\mid z \mid = \sqrt{x^2 + y^2}$

and $$\left| \frac{w_n}{w^{n-1}} \right| = \frac{\mid z \mid}{n} \to 0 \text{ as } n \to \infty \quad \text{for all } z.$$

Hence, by d'Alembert's Test of 11.4, $\Sigma \mid w_n \mid$ is convergent, and the power series (by the definition) is absolutely convergent for all z.

As a matter of notation, the sum of the power series is a function of a complex variable written e^z:

NOTATION: $e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \ldots$, *absolutely convergent for any complex* $z = x + iy$.

This is an extension of e^x, and e^z reduces to e^x when $y=0$. The basic property is obtained by multiplication exactly as in 12.2:

$$e^{z_1} \times e^{z_2} = e^{z_1+z_2} \quad \text{for all } z_1 \text{ and } z_2 \quad \ldots\ldots\ldots\ldots\ldots\ldots(2)$$

As a case of (2), take $z_1 = x$ and $z_2 = iy$ (x and y real): $e^x \times e^{iy} = e^{x+iy}$. Hence, if $z = x + iy$ is any complex number, e^z is the product of a real factor e^x and a complex value e^{iy}:

$$e^z = e^x \times e^{iy} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

To complete the form (3), an explicit expression of e^{iy} as a complex number is required. From the power series:

$$e^{iy} = 1 + (iy) + \frac{(iy)^2}{2!} + \frac{(iy)^3}{3!} + \ldots \quad (i^2 = -1)$$

$$= \left(1 - \frac{y^2}{2!} + \frac{y^4}{4!} - \ldots\right) + i\left(y - \frac{y^3}{3!} + \frac{y^5}{5!} - \ldots\right)$$

i.e.
$$e^{iy} = \cos y + i \sin y \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

by the power series for the circular functions. Putting (4) into (3):

$$e^z = e^x(\cos y + i \sin y) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

So, if $Z = e^z$ and $Z = X + iY$, then $X = e^x \cos y$ and $Y = e^x \sin y$. The expression (5) is all we need; the conformal transformation $Z = e^z$ is very simple:

$$X = e^x \cos y \quad \text{and} \quad Y = e^x \sin y.$$

Notice that the definition of $\cos x$ and $\sin x$ gives:

$$\cos(-x) = \cos x \quad \text{and} \quad \sin(-x) = -\sin x$$

which is sometimes summarised by saying that $\cos x$ is an 'even' function and that $\sin x$ is an 'odd' function. Hence (4) gives:

$$\cos x + i \sin x = e^{ix} \quad \text{and} \quad \cos x - i \sin x = e^{-ix} \quad \ldots\ldots\ldots(6)$$

Add (6) to give: $\quad 2\cos x = e^{ix} + e^{-ix}$

and subtract: $\quad 2i \sin x = e^{ix} - e^{-ix}$.

So: $$\cos x = \frac{1}{2}\left(e^{ix} + e^{-ix}\right) \quad \text{and} \quad \sin x = \frac{1}{2i}\left(e^{ix} - e^{-ix}\right). \ldots\ldots\ldots\ldots(7)$$

The results (7) are remarkable, and most useful. The expressions on the right-hand sides are complex values; but in each case the 'imaginary' part disappears and the expressions reduce to real values ($\cos x$ and $\sin x$ respectively). They also suggest a further notation, in real values only:

NOTATION: *cosh* $x = \frac{1}{2}(e^x + e^{-x})$ *and* *sinh* $x = \frac{1}{2}(e^x - e^{-x})$.

Cosh x and sinh x are called *hyperbolic functions*. There is nothing new about them; they are merely convenient notations for the exponential expressions shown.

12.7. Trigonometric functions. The circular functions, as defined above, have nothing whatever to do with trigonometry. For any real value x, $\cos x$ and $\sin x$ are obtained as the sums of particular series and $\tan x$ is $\sin x/\cos x$. It is *not* assumed that x is the measure of an angle; it is *not* assumed that $\cos x$, $\sin x$ and $\tan x$ are trigonometric ratios between the sides of a right-angled triangle.

We now proceed to establish that they can be interpreted in these ways. First, what do we mean by the measure of an angle? We may try to pass the buck by saying that the measure is such that, if an arc AP of a circle (centre O, radius r) subtends an angle θ at O, the length of the arc AP is $r\theta$ and the area of the sector AOP is $\frac{1}{2}r^2\theta$. But the pass must be refused. The question remains: what do we mean by the length of an arc or the area of a sector? Nothing is defined; we need to remedy the defect. We choose to do so by applying the general concept of an area (10.4) to a circle.* At the same time, we introduce the trigonometric ratios and link them with the circular functions.

To put the problem specifically: in Fig. 12.7a, let A be a given point and P a variable point on a circle, centre O and radius r. A measure x is to be assigned to the angle AOP in such a way that the area of the sector AOP is $\frac{1}{2}r^2x$. Further, with this measure x, and with $\tan x$ as defined in 12.5, $\tan x$ is to

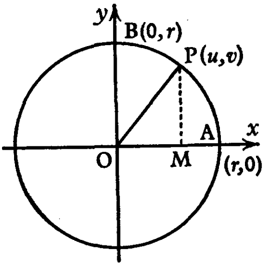


FIG. 12.7a

* The length of an arc of a curve can also be defined in terms of integrals, and it can then be applied to a circle. This is more difficult than for areas and we are well advised to stick to areas here.

be shown equal to the trigonometric ratio of MP to OM, where PM is perpendicular to OA. This is a formidable assignment.

Insert axes Ox and Oy as shown in Fig. 12.7a and take (for the moment) the point $P\ (u, v)$ in the positive quadrant, $\angle AOP$ between zero and a right angle. Write $t=MP/OM$, the ratio with which we are most concerned. Any point $P\ (u, v)$ on the quarter circle from $A\ (r, 0)$ to $B\ (0, r)$ satisfies $u^2+v^2=r^2$ $(u>0,\ v>0)$. Now $t=\frac{v}{u}$ so that $u^2+t^2u^2=r^2$ $(u>0,\ t>0)$. Hence:

$$u=r/\surd(1+t^2) \quad \text{and} \quad v=rt/\surd(1+t^2) \quad (t>0).$$

It follows that the point (x, y), where $x=\frac{r}{\sqrt{1+\tau^2}}$ and $y=\frac{r\tau}{\sqrt{1+\tau^2}}$, describes the arc AP from S to P as τ increases from O to t. The area PMA under the circle from M to A is given by:

$$\int_{x=u}^{r} y\,dx=\int_{\tau=t}^{0} y\,\frac{dx}{d\tau}\,d\tau=-\int_0^t \frac{r\tau}{\sqrt{1+\tau^2}}\,\frac{-r\tau}{\sqrt{(1+\tau^2)^3}}\,d\tau$$

$$=r^2\int_0^t \frac{\tau^2\,d\tau}{(1+\tau^2)^2}=r^2\int_0^t \tau\times\frac{\tau\,d\tau}{(1+\tau^2)^2}$$

$$=r^2\left[\tau\,\frac{-1}{2(1+\tau^2)}+\int\frac{d\tau}{2(1+\tau^2)}\right]_0^t$$

on integrating by parts. Hence the area PMA is:

$$-\tfrac{1}{2}\frac{r^2t}{1+t^2}+\tfrac{1}{2}r^2\tan^{-1}t.$$

The area of the triangle OPM is $\frac{1}{2}uv=\frac{1}{2}\frac{r^2t}{1+t^2}$. The area S of the sector AOP is the sum of the area PMA and the triangle OPM. Hence:

$$S=\tfrac{1}{2}r^2\tan^{-1}t.$$

We now have our answers. As a definition, take the measure of the angle AOP as $x=\tan^{-1}t$ and call the unit radians. Here the variable t is the trigonometric ratio of MP to OM. Hence $t=\tan x$, so that $\tan x$ (x measure of the angle in radians) is the trigonometric ratio t. Further, the area of the sector AOP is $\frac{1}{2}r^2x$.

DEFINITION: *The measure of the angle AOP in radians is* $x=\tan^{-1}t$ *where t is the trigonometric ratio MP/OM.*

This is for $t>0$, i.e. P in the positive quadrant. The two extreme values, at A and B, have $t=0$ and $t \to \infty$ respectively.

The constant π can now be introduced: $\tan \frac{1}{4}\pi = 1$ by (9) of 12.5, i.e. $\pi/4$ is the angle for which $t = MP/OM = 1$. The triangle OMP is then isosceles, half a square, i.e. the angle $\pi/4$ is half a right angle. So:

$$\text{Right angle} = \tfrac{1}{2}\pi \text{ radians}$$

and at A $(t=0)$ the angle $x=0$, at B $(t \to \infty)$ the angle $x = \frac{1}{2}\pi$ radians.* Since the area of the sector AOP is $\frac{1}{2}r^2x$, it follows that the area of the quarter-circle $(x=\frac{1}{2}\pi)$ is $\frac{1}{4}r^2\pi$ and the area of the circle is πr^2.

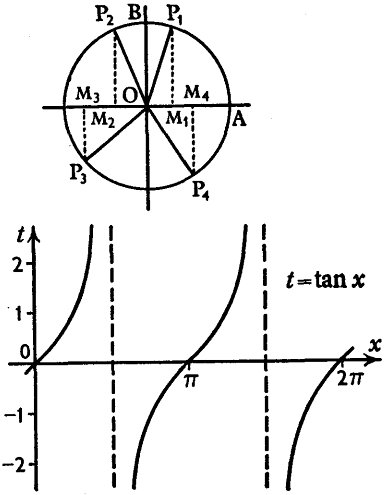


FIG. 12.7*b*

If t is the trigonometric ratio MP/OM, then the function $t = \tan x$ increases over the range $t \geqslant 0$ as x increases over the domain

$$0 \leqslant x \leqslant \tfrac{1}{2}\pi.$$

The domain of $t = \tan x$ can be extended, by letting P swing round the circle, anti-clockwise as in Fig. 12.7*b*, running through four right angles. Then:

Angle x	*Sign of OM*	*Sign of MP*	*t = tan x Sign*	*Range*
$0 \leqslant x \leqslant \frac{\pi}{2}$	+	+	+	0 to ∞
$\frac{\pi}{2} \leqslant x \leqslant \pi$	−	+	−	− ∞ to 0
$\pi \leqslant x \leqslant \frac{3\pi}{2}$	−	−	+	0 to ∞
$\frac{3\pi}{2} \leqslant x \leqslant 2\pi$	+	−	−	− ∞ to 0

The graph of $t = \tan x$, over the domain $0 \leqslant x \leqslant 2\pi$, is as shown. For larger x, as P swings around the circle for a second, third, ... time,

* To put into degrees (the elementary angle measure), we re-scale so that the right angle is 90°: $\frac{1}{2}\pi$ radians $= 90°$. An angle of x radians $= \dfrac{180}{\pi}\, x°$.

the values of $t = \tan x$ repeat themselves. Similarly, $t = \tan x$ repeats to the left, for negative values of x, P swinging around the circle in the clockwise direction.

Finally, having got $\tan x$ as the trigonometric ratio MP/OM (in Fig. 12.7a), we proceed to interpret $\cos x$ and $\sin x$ similarly. If $t = \tan x$, then

$$\sin x/\cos x = t \quad \text{and} \quad \sin^2 x + \cos^2 x = 1.$$

Hence: $\sin x = t \cos x$ and $t^2 \cos^2 x + \cos^2 x = 1$

i.e. $\cos x = 1/\surd(1+t^2)$ and $\sin x = t/\surd(1+t^2)$.

Since $t = \dfrac{MP}{OM}$, $1 + t^2 = \dfrac{OM^2 + MP^2}{OM^2} = \dfrac{OP^2}{OM^2}$

i.e. $\cos x = OM/OP$ and $\sin x = MP/OP$

which are the well-known trigonometric ratios. This is for $0 \leqslant x \leqslant \dfrac{\pi}{2}$ but the extension to other x follows as for $\tan x$. Referring to Fig. 12.7b, we see that, as x increases from 0 (at A), $\cos x$ decreases from 1 ($x = 0$) to 0 ($x = \frac{1}{2}\pi$), to -1 ($x = \pi$); and then increases again to 0 $\left(x = \dfrac{3\pi}{2}\right)$ and back to 1 ($x = 2\pi$). This cycle is repeated in subsequent intervals $2\pi \leqslant x \leqslant 4\pi$, $4\pi \leqslant x \leqslant 6\pi$, ... and similarly for $x < 0$. The graph of $y = \cos x$ is shown in Fig. 12.7c. The graph of $y = \sin x$ is similar, but $\sin x = 0$ at $x = 0$ and $\sin x = 1$ at $x = \frac{1}{2}\pi$. Both functions are *periodic*, repeating themselves in periods of 2π for x.

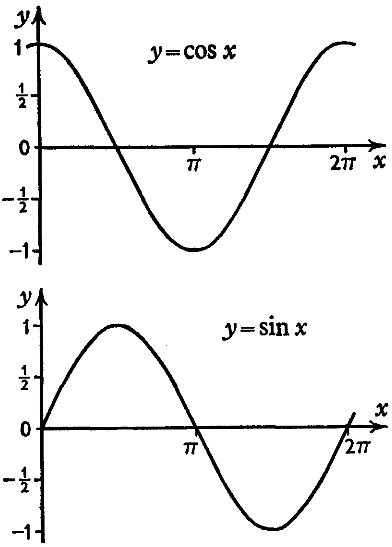


FIG. 12.7c

Consider now the representation of a complex number $z = x + iy$ as a point P (x, y) on an Argand Diagram, as in Fig. 2.5a above. Let $OP = r$ and let the angle OP makes with Ox be θ radians. Then:

$$x/r = OM/OP = \cos\theta$$

i.e. $x = r\cos\theta$

and $y/r = MP/OP = \sin\theta$

i.e. $y = r\sin\theta$.

Hence: $z = x + iy = r(\cos\theta + i\sin\theta)$.

Further: $x^2 + y^2 = r^2(\cos^2\theta + \sin^2\theta) = r^2$

and $\dfrac{y}{x} = \dfrac{r\cos\theta}{r\sin\theta} = \tan\theta$.

Hence: $r = \surd(x^2 + y^2)$ and $\theta = \tan^{-1} y/x$

where $r = |z|$ is the *absolute value* or modulus of the complex number z and where θ is its *argument* or amplitude (subject to the condition of Appendix A.9).

The results (6) or (7) of 12.6 are useful in the further development of a complex number in terms of its absolute value r and argument θ:

$$z = r(\cos\theta + i\sin\theta) = re^{i\theta} \quad\text{.......................}(1)$$

Multiplication is easy with (1). If $z_1 = r_1e^{i\theta_1}$ and $z_2 = r_2e^{i\theta_2}$:

$$z = z_1z_2 = r_1e^{i\theta_1} \times r_2e^{i\theta_2} = r_1r_2e^{i(\theta_1+\theta_2)} = re^{i\theta}$$

where $r = r_1r_2$ is the absolute value and $\theta = \theta_1 + \theta_2$ is the argument of z. In a product of complex numbers, absolute values are multiplied and arguments added.

The remarkable relations between the basic constants e and π can now be given. Put $\theta = \frac{1}{2}\pi$ in (1), noting that $\cos\frac{1}{2}\pi = 0$ and $\sin\frac{1}{2}\pi = 1$: $e^{i\pi/2} = i$. Similar results follow when $\theta = \pi, \dfrac{3\pi}{2}, 2\pi$ are substituted:

$$e^{i\pi/2} = i;\ e^{\pi i} = -1;\ e^{3i\pi/2} = -i;\ e^{2\pi i} = 1 \quad\text{.................}(2)$$

The relations (2) can be interpreted in a variety of ways. They express the fact that 'multiplication by i' means 'rotation through a right-angle' on an Argand Diagram.* They represent in turn the fixed points B, A', B' and A on the Diagram (Fig. 2.5*a*). For, B is (0, 1) in Cartesian co-ordinates, i.e. $z = 0 + 1i = i$; the absolute value of z is $r = 1$ and argument $\theta = \frac{1}{2}\pi$, i.e. $z = 1 \times e^{i\pi/2} = e^{i\pi/2}$. The others follow similarly. Finally, the set $\{i, -1, -i, 1\}$ form a cyclic group, and so does the set

$$\{e^{i\pi/2}, e^{\pi i}, e^{3i\pi/2}, e^{2\pi i}\}.$$

The fact that this is a cyclic group follows from the relation: $e^{2\pi i} = 1$.

* Multiply $z = re^{i\theta}$ by $i = e^{i\pi/2}$: $z \times i = re^{i\theta}e^{i\pi/2} = re^{i(\theta+\pi/2)}$. The complex number z is rotated through a right angle on an Argand Diagram.

12.8. Summary of results

Properties:

$$e^{z_1} \times e^{z_2} = e^{z_1+z_2} \qquad a^x = e^{x \log a}$$

$$\log xy = \log x + \log y \qquad \log x^b = b \log x$$

$$\sin^2 x + \cos^2 x = 1 \qquad \tan x = \sin x / \cos x$$

$$\cos(x+y) = \cos x \cos y - \sin x \sin y \qquad \sin(x+y) = \sin x \cos y + \cos x \sin y$$

$$\cos x + i \sin x = e^{ix} \qquad \cos x - i \sin x = e^{-ix}$$

$$\cos x = \tfrac{1}{2}(e^{ix} + e^{-ix}) \qquad \sin x = \frac{1}{2i}\left(e^{ix} - e^{-ix}\right)$$

$$z = x + iy = r(\cos\theta + i \sin\theta) = re^{i\theta}$$

where $r = \sqrt{x^2+y^2}$ (absolute value), $\theta = \tan^{-1} y/x$ (argument)

$$e^{2\pi i} = 1.$$

Expansions:

$$e^x = \exp x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots \qquad \text{all } x$$

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots \qquad -1 < x \leqslant 1$$

$$(1+x)^a = 1 + ax + \frac{a(a-1)}{2!}x^2 + \frac{a(a-1)(a-2)}{3!}x^3 + \dots \qquad -1 < x < 1$$

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots \qquad \text{all } x$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \dots \qquad \text{all } x$$

$$\tan^{-1} x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \dots \qquad -1 \leqslant x \leqslant 1$$

Derivatives:

$$De^x = e^x;\ De^{f(x)} = f'(x)e^{f(x)} \qquad Da^x = a^x \log a$$

$$D \log x = \frac{1}{x};\ D \log f(x) = \frac{f'(x)}{f(x)} \qquad Dx^a = ax^{a-1}$$

$$D \cos x = -\sin x \qquad D \sin x = \cos x$$

$$D \tan x = 1 + \tan^2 x \qquad D \tan^{-1} x = 1/1 + x^2$$

Integrals:

$$\int e^x\,dx = e^x;\quad \int e^{f(x)}f'(x)\,dx = e^{f(x)} \qquad \int a^x\,dx = \frac{a^x}{\log a}$$

$$\int \frac{dx}{x} = \log x;\quad \int \frac{f'(x)}{f(x)}\,dx = \log f(x) \qquad \int x^a\,dx = \frac{x^{a+1}}{a+1} \quad (a \neq -1)$$

$$\int \cos x\,dx = \sin x \qquad \int \sin x\,dx = -\cos x$$

$$\int \frac{dx}{1+x^2} = \tan^{-1} x$$

Limits:

$$\operatorname*{Lim}_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x \qquad \operatorname*{Lim}_{x\to 0}\frac{\sin x}{x} = 1$$

$$\operatorname*{Lim}_{x\to\infty}\frac{\log x}{x^a} = \operatorname*{Lim}_{x\to\infty}\frac{x^a}{e^x} = 0 \quad (a>0)$$

12.9. Exercises

1. Show that the derivative and integral of e^{-x} are both $-e^{-x}$. Write $De^{\frac{1}{2}x^2}$ and $De^{-\frac{1}{2}x^2}$ and deduce that $\int xe^{\frac{1}{2}x^2}\,dx = e^{\frac{1}{2}x^2}$ and $\int xe^{-\frac{1}{2}x^2}\,dx = -e^{-\frac{1}{2}x^2}$.

2. *Normal distribution.* Show that $y = y_0e^{-\frac{1}{2}x^2}$ has a single extreme value, a maximum y_0 at $x=0$, and that $y \to 0$ as $x \to \pm\infty$. Indicate the shape of the curve $y = y_0e^{-\frac{1}{2}x^2}$, the basic form of the 'normal distribution' of statistics.

3. By the standard form (3) of 12.3 show that $D\log e^x = 1$ and deduce that $\log e^x = x$.

4. Write $D\log(1+x)$ and $D\log(1+x^2)$ and generalise to:

$$D\log(a_nx^n + a_{n-1}x^{n-1} + \ldots + a_1x + a_0) = \frac{na_nx^{n-1} + (n-1)a_{n-1}x^{n-2} + \ldots + a_1}{a_nx^n + a_{n-1}x^{n-1} + \ldots + a_1x + a_0}.$$

5. Integrate by parts to derive: $\int xe^{-x}\,dx = -xe^{-x} + \int e^{-x}\,dx = -e^{-x}(x+1)$. Use $xe^{-x} \to 0$ as $x \to \infty$ to deduce that $\int_0^\infty xe^{-x}\,dx = 1$ is convergent.

6. By the method and result of Ex. 5, show that $I_n = \int x^ne^{-x}\,dx$ can be expressed by the *reduction formula*: $I_n = -x^ne^{-x} + nI_{n-1}$; and that $\int_0^\infty x^ne^{-x}\,dx$ is convergent. Deduce that $\int_0^\infty x^ne^{-x}\,dx = n\int_0^\infty x^{n-1}e^{-x}\,dx$ and that n factoria can be written as this infinite integral: $n! = \int_0^\infty x^ne^{-x}\,dx$.

7. *Algebraic and non-algebraic functions.* Establish the following:

function	*derivative*	*integral*
$1+x$	1	$\frac{1}{2}(1+x)^2$
$\dfrac{1}{1+x}$	$-\dfrac{1}{(1+x)^2}$	$\log(1+x) \quad (x>-1)$
$\dfrac{1}{1-x}$	$\dfrac{1}{(1-x)^2}$	$\log\left(\dfrac{1}{1-x}\right) \quad (x<1)$
$\dfrac{1}{1+x^2}$	$-\dfrac{2x}{(1+x^2)^2}$	$\tan^{-1} x$
$\dfrac{1}{1-x^2}$	$\dfrac{2x}{(1-x^2)^2}$	$\frac{1}{2}\log\left(\dfrac{1+x}{1-x}\right) \quad (x<1)$

Hence illustrate the fact that derivatives of algebraic functions *must* be algebraic and that derivatives of non-algebraic functions *may* be algebraic.

8. Check back, by writing derivatives, that:
$\int\frac{dx}{\sqrt{x(x-1)}}=2\log(\sqrt{x}+\sqrt{x-1})$ (for $x>1$); $=-2\log(\sqrt{-x}+\sqrt{1-x})$ (for $x<0$), with $\frac{1}{\sqrt{x(x-1)}}$ not defined for $0<x<1$. Show that

$$\int\frac{dx}{\sqrt{x(1-x)}}=2\tan^{-1}\sqrt{\frac{x}{1-x}} \text{ for } 0<x<1.$$

Draw graphs of $y=\frac{1}{\sqrt{x(x-1)}}$ and of $y=\frac{1}{\sqrt{x(1-x)}}$ to illustrate.

9. Given $\frac{1}{y}Dy=D\log y=r$ (constant), show that $\log y=rx+A$ (A constant) and that the only function to satisfy the given condition is the exponential $y=y_0e^{rx}$. (Here $y_0=\log A$.) Can r be taken as negative as well as positive?

*10. A function $f(x)$ satisfies the relation $f(x)\times f(y)=f(x+y)$ for all real x and y. Show that the exponential function $y=y_0e^{rx}$ is *one* function which does. To prove the converse, that it is the *only* function which does, proceed as follows. If $f(x)$ satisfies the relation, show that $f'(x)=\frac{f'(n)}{f(y)}$ where $n=x+y$, y fixed. Similarly, $f'(y)=\frac{f'(n)}{f(x)}$. Deduce that $\frac{f'(x)}{f(x)}=\frac{f'(y)}{f(y)}=$constant ($r$), all x and y; and so that $f'(x)=rf(x)$. Use Ex. 9 to derive the required result.

11. From the definition of exp x and (1) of 12.2, show that $\frac{1}{h}(\exp h-1)\rightarrow 1$ as $h\rightarrow 0$ and that $\frac{1}{h}\{\exp(x+h)-\exp x\}\rightarrow\exp x$ as $h\rightarrow 0$. This establishes from first principles that exp x is continuous with $D\exp x=\exp x$.

12. *Semi-logarithmic graphs.* The rate of change $f'(\alpha)$ is the tangent slope of the graph of $y=f(x)$ at the point where $x=\alpha$. Show that the proportionate rate of change of $f(x)$ at $x=\alpha$ is the tangent slope of the graph of $u=\log f(x)$

at the point $x=\alpha$. The graph of $u=\log f(x)$ is called the semi-logarithmic graph of $y=f(x)$; it plots $\log y$ rather than y against x. Show that the exponential function $y=y_0e^{rx}$ is a line of slope r on a semi-logarithmic graph.

13. *Compound interest.* Let £x be the amount of £a after t years when interest is compounded at $100r$ per cent per year. Show that $x=a(1+r)^t$ for yearly compounding and $x=a(1+\frac{1}{2}r)^{2t}$ for twice-yearly compounding. Generally, reckoning interest n times a year, show that $x=a\left(1+\frac{r}{n}\right)^{nt}$. Let $n\to\infty$ and use (6) of 12.3 to show that the exponential $x=ae^{rt}$ represents growth when interest is compounded continuously at the rate of $100r$ per cent per year.

14. An investment doubles in n years at interest compounded annually at ρ per cent per year. Show that n is a function of ρ given by $\left(1+\frac{\rho}{100}\right)^n=2$. Use logarithm tables to show that $n=14{\cdot}2$ approximately when $\rho=5$. Generally show that

$$\frac{\log 2}{n}=\frac{\rho}{100}-\frac{1}{2}\left(\frac{\rho}{100}\right)^2+\frac{1}{3}\left(\frac{\rho}{100}\right)^3-\dots$$

Neglect ρ^2, take the natural logarithm $\log 2=0{\cdot}7$ approximately, and get the practical rule of thumb: $n=\frac{70}{\rho}$ approximately (e.g. $n=\frac{70}{5}=14$ for $\rho=5$).

15. *Population growth.* Let x be the number of births at time t. Assume that x increases at the steady proportionate rate of $100r$ per cent per year and that the life span is 50 years (everyone dies aged 50). Write $y=\int_{t-50}^{t}x\,dt$ for the total live population, and $u=\int_{-\infty}^{t}x\,dt$ for the number ever born, at time t. Show that y and u increase at $100r$ per cent per year and that $\frac{y}{u}=1-e^{-50r}=$ constant over time. For a 2 per cent increase ($r=0{\cdot}02$) show that the live population is always 63 per cent of the total population throughout time, i.e. that the present population outnumbers their ancestors.

16. Show that $\log_{10}(1+x)=0{\cdot}43429\dots(x-\frac{1}{2}x^2+\frac{1}{3}x^3-\frac{1}{4}x^4+\dots)$ for $|x|<1$ and for $x=1$. Indicate how this gives a method of evaluating $\log_{10}a$, directly for $1\leqslant a\leqslant 2$ and by use of (4) of 12.4 for other a.

17. Write $De^{a\log x}=\frac{a}{x}e^{a\log x}$ and deduce $Dx^a=ax^{a-1}$ and $\int x^a\,dx=\frac{x^{a+1}}{a+1}$ ($a\neq-1$). From successive derivatives of x^a at $x=1$, obtain the expansion of $(1+x)^a$ for real a as a Taylor's series.

18. Graph $y=x^3$ and $y=e^x-1$ for $x>0$, using the same axes and scales and illustrate that x^3 increases less quickly than e^x as $x\to\infty$. Similarly illustrate the tendencies of $\sqrt{x}$ and $\log x$ as $x\to\infty$ by a graph of $y=\sqrt{x}$ and $y=1+\log x$.

*19. Write the general term of the series for $\cos x$ as $(-1)^n\frac{x^{2n}}{(2n!)}$ and show that the general term of the product of the series for $\cos x$ and $\cos y$ is

$$\frac{(-1)^n}{(2n)!}\left\{x^{2n}+\frac{2n(2n-1)}{2!}x^{2n-2}y^2+\frac{2n(2n-1)(2n-2)(2n-3)}{4!}x^{2n-4}y^4+\ldots+y^{2n}\right\}.$$

Write the corresponding term in the product of $\sin x$ and $\sin y$ and deduce that the general term in the expansion of $\cos x \cos y - \sin x \sin y$ is

$$\frac{(-1)^n}{(2n)!}\left\{x^{2n}+2nx^{2n-1}y+\frac{2n(2n-1)}{2!}x^{2n-2}y^2+\ldots+2nxy^{2n-1}+y^{2n}\right\}.$$

Identify the terms in brackets as $(x+y)^{2n}$ and the expansion as that of $\cos(x+y)$.

*20. Repeat the steps of Ex. 19 to show that the general term in the expansion of $\sin x \cos y + \cos x \sin y$ is the same as that in $\sin(x+y)$.

21. Obtain $\int e^x \cos x\, dx = e^x \cos x + \int e^x \sin x\, dx$ and a similar result for $\int e^x \sin x\, dx$ by integration by parts. Deduce that

$$\int e^x \cos x\, dx = \tfrac{1}{2}e^x(\sin x + \cos x) \quad \text{and} \quad \int e^x \sin x\, dx = \tfrac{1}{2}e^x(\sin x - \cos x).$$

22. By the method of Ex. 21, find $\int e^{-x} \cos x\, dx$ and $\int e^{-x} \sin x\, dx$ and show that

$$\int_0^\infty e^{-x}\cos x\, dx = \int_0^\infty e^{-x} \sin x\, dx = \tfrac{1}{2}.$$

*23. In view of the fact that $\sin x$ is bounded (oscillatory with period 2π), what can be said about the limits of $\sin x$ as $x \to \infty$ and of $\sin\left(\frac{1}{x}\right)$ as $x \to 0$? If $f(x) = \sin x\pi$, show that it is fallacious to infer $\lim\limits_{x\to\infty} f(x)$ from the limit of the sequence $f(n)$ for $n = 1, 2, 3, \ldots$. (In this case $f(n) = 0$ all n.) Consider the crescendo oscillation of $\sin\left(\frac{1}{x}\right)$ for small x and indicate the nature of the discontinuity of $\sin\left(\frac{1}{x}\right)$ at $x = 0$.

24. *Hyperbolic functions.* From the definition (12.6), show that $\cosh x$ is an 'even' function and $\sinh x$ an 'odd' function such that $\cosh^2 x - \sinh^2 x = 1$. Show that $D \cosh x = \sinh x$ and $D \sinh x = \cosh x$; check that $D \tan^{-1}(e^x)$ can be expressed as the reciprocal of $2 \cosh x$.

25. Deduce from $\tan \frac{1}{4}\pi = 1$ that $\cos \frac{1}{4}\pi = \sin \frac{1}{4}\pi = \frac{1}{\sqrt{2}}$ and check by the trigonometric ratios of 45° ($\frac{1}{4}\pi$ radians) in an isosceles right-angled triangle.

26. A cube root of unity is $\omega = \frac{1}{2}(-1 + i\sqrt{3})$. By the corresponding point on an Argand diagram (3.8 above), show that the argument of ω is $\frac{2\pi}{3}$ radians. Deduce from (1) of 12.7 that $\cos \frac{2\pi}{3} = -\frac{1}{2}$ and $\sin \frac{2\pi}{3} = \frac{\sqrt{3}}{2}$.

27. From the addition formula for $\cos(x+y)$ show that

$$\cos 2x = 2\cos^2 x - 1 = 1 - 2\sin^2 x.$$

Hence show that $\cos \frac{2\pi}{3} = -\frac{1}{2}$ gives $\cos \frac{1}{3}\pi = \frac{1}{2}$ and $\sin \frac{1}{3}\pi = \frac{\sqrt{3}}{2}$. Check from the trigonometric ratios of 60° ($\frac{1}{3}\pi$ radians) in an equilateral triangle.

*28. *Beta functions.* Take the definition $B(m, n) = \int_0^1 x^{m-1}(1-x)^{n-1}\, dx$ for any real $m>0$, $n>0$. Substitute $x=\sin^2\theta$ to transform $B(m, n)$ into $2\int_0^{\frac{1}{2}\pi}\sin^{2m-1}\theta\cos^{2n-1}\theta\,d\theta$ and deduce that $B(\frac{1}{2},\frac{1}{2})=\pi$. By this and the other transform of $B(m, n)$ given in 10.9 Ex. 27 show that π can be expressed as

$$\int_0^1\frac{dx}{\sqrt{x(1-x)}}=\int_0^\infty\frac{dx}{(1+x)\sqrt{x}}\,.$$

Check the first from Ex. 8 above and the second by substituting $x=t^2$ in the integral.

*29. *Gamma functions.* Define $\Gamma(n)=\int_0^\infty x^{n-1}e^{-x}\,dx$ for any real $n>0$. Show by the method of Ex. 6 above that $\Gamma(n)=(n-1)\Gamma(n-1)$. If n is integral, deduce that $\Gamma(1)=1$ and $\Gamma(n)=(n-1)!$

*30. Substitute $x=\frac{1}{2}t^2$ in $\Gamma(n)$ and show that $\Gamma(n)=\frac{1}{2^{n-1}}\int_0^\infty t^{2n-1}e^{-\frac{1}{2}t^2}\,dt$.

Given that the 'normal distribution' $y=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$ has unit area, check that $\int_0^\infty y\,dx=\frac{1}{2}$, and that $\Gamma(\frac{1}{2})=\sqrt{\pi}$.

*31. *Hypergeometric series.* Write $F(\alpha, \beta; \gamma; x)$ for the sum of the hypergeometric series of 11.9 Ex. 29, absolutely convergent for $|x|<1$. Show that several familiar series are particular cases: Binomial $(1-x)^{-m}=F(m, 1; 1; x)$; Logarithmic $\log(1+x)=xF(1, 1; 2; -x)$; Inverse tangent $\tan^{-1}x=xF(\frac{1}{2}, 1; \frac{3}{2}; -x^2)$.

*32. Write the hypergeometric series $F(\alpha, 1; 1; x/\alpha)$ for $|x|<\alpha$. Show that the exponential series for e^x can be regarded as the limit as $\alpha\to\infty$ of $F(\alpha, 1; 1; x/\alpha)$.

CHAPTER 13

LINEAR ALGEBRA

13.1. The basis of linear algebra. The concept of a vector space, though defined in purely algebraic terms, was used in Chapter 8 as a basis for geometric space and, in particular, for Euclidean space. The object now is to explore the purely algebraic properties and applications of vector spaces, the emphasis being on algebraic vectors.

A set of vectors over a field of scalars is the foundation of linear algebra, a vast subject with many applications. The idea of 'linearity' is present in the definition of a vector space: a set V of entities called vectors, and subject to the operation of addition, together with an outside set F of scalars used in the operation of scalar multiplication. The result is that, given vectors $v_1, v_2, \ldots$ of V and scalars $a_1, a_2, \ldots$ of F, then $a_1v_1 + a_2v_2 + \ldots$ is another vector of V. Hence, sums and scalar products together serve to give the familiar 'linear' form $a_1v_1 + a_2v_2 + \ldots$ in a general setting: the algebraic sum of multiples of certain vectors. The following simple cases indicate how varied is the interpretation of the 'linear' form.

(i) Take V as the field of real numbers and F as the field of rationals. A typical vector is the real variable x and a typical scalar is the rational multiple a. The linear form is $a_1x_1 + a_2x_2 + \ldots$, also a real number of V. So real numbers form a vector space over the field of rationals. This vector space handles linear forms of real variables with rational coefficients.

(ii) Take V as the set of real number pairs (x, y) and F as the field of real numbers. Add pairs by the rule:

$$(x_1, y_1) + (x_2, y_2) = (x_1 + x_2, y_1 + y_2).$$

Define scalar products by the rule: $a(x, y) = (ax, ay)$. V is then a vector space over F and it deals with linear combinations of the form:

$$a_1(x_1, y_1) + a_2(x_2, y_2) + \ldots = (a_1x_1 + a_2x_2 + \ldots, a_1y_1 + a_2y_2 + \ldots).$$

This is, in fact, the primitive notion of a vector as a point P, or a directed line OP, in a plane. For example, the mid-point of $P_1(x_1, y_1)$ and $P_2(x_2, y_2)$ is the linear combination:

$$\tfrac{1}{2}(x_1, y_1)+\tfrac{1}{2}(x_2, y_2)=\{\tfrac{1}{2}(x_1+x_2), \tfrac{1}{2}(y_1+y_2)\}.$$

Another interpretation of the same vector space is the field of complex numbers, $z=(x, y)=x+iy$, over the field of real numbers. The linear form is then:

$$a_1z_1+a_2z_2+\ldots=(a_1x_1+a_2x_2+\ldots)+i(a_1y_1+a_2y_2+\ldots).$$

The representation of a complex number as a point on an Argand diagram (2.5) is the link between the two aspects of the vector space.

(iii) Take V as the set of quadratic polynomials with rational (or real) coefficients and F as the field of rational (or real) scalars. The typical vector is then the triple (a, b, c) of elements of F or the expression ax^2+bx+c in the undefined x. Adding and multiplying quadratics by scalar multiples in the ordinary way, we get V as a vector space. The linear combination of quadratics, with λ's from F, is:

$$\begin{aligned}&\lambda_1(a_1x^2+b_1x+c_1)+\lambda_2(a_2x^2+b_2x+c_2)+\ldots\\&\qquad=(\lambda_1a_1+\lambda_2a_2+\ldots)x^2+(\lambda_1b_1+\lambda_2b_2+\ldots)x+(\lambda_1c_1+\lambda_2c_2+\ldots).\end{aligned}$$

The development of linear algebra is on the following lines. One vector v of a vector space V is a *linear combination* of other vectors $v_1, v_2, \ldots$ if it can be written: $v=a_1v_1+a_2v_2+\ldots$ for some scalars $a_1, a_2, \ldots$. The concept of a *linear transformation*, as in 7.5, can then be made perfectly general: a mapping of one vector space V into another vector space V' such that a linear combination of vectors of V is mapped into the *same* linear combination of vectors of V'.

The general vector space V over F can be specialised to a more practical form by taking vectors as n-tuples $v=(x_1, x_2, \ldots x_n)$ of values from the field F (usually real numbers) which also provides the scalars for scalar products of the n-tuples. The space $V_n(F)$ of n-tuples is then obtained and Euclidean space $E_n(F)$ is a particular case (8.4). The algebraic concept of a space of n-tuples is of very wide scope; example (ii) here is a case of $V_2(F)$ and example (iii) of $V_3(F)$. The basic result, proved in 15.9, is that, apart from vector spaces of special form (of 'infinite' dimension), any vector space V whatever is isomorphic with, and algebraically indistinguishable from, a space of n-tuples for some integral n, the dimension of V.

The general concept of a linear transformation becomes specialised, and more familiar, when applied as a mapping of the space $V_n(F)$ of n-tuples into the space $V_m(F)$ of m-tuples. If the n-tuple $(x_1, x_2, \dots x_n)$ maps into the m-tuple $(y_1, y_2, \dots y_m)$, then y_1 is a linear expression in the n real variables $x_1, x_2, \dots x_n$, and similarly for $y_2, y_3, \dots y_m$. For example, as in 7.5, if $n=m=2$, then $y_1=a_{11}x_1+a_{12}x_2$ and $y_2=a_{21}x_1+a_{22}x_2$ is the linear transformation, completely described by the double array

$$\begin{Vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{Vmatrix}$$

of scalar values. This is the concept of a *matrix*, of two rows and columns in the particular case $n=m=2$, and generally of m rows and n columns. The matrix notation is introduced initially to lighten the algebraic burden of linear transformations and equations; it is later found to have a great variety of other applications.

Consider the problem of 'inverting' a linear transformation T from the space of n-tuples to that of m-tuples. T appears as linear expressions for $y_1, y_2, \dots y_m$ in terms of the variables $x_1, x_2, \dots x_n$. Can we turn T around so that it gives expressions for $x_1, x_2, \dots x_n$ in terms of the variables $y_1, y_{,2} \dots y_m$? T is arranged to provide values of the y's when values are assigned to the x's; can T also turn the trick the other way? Exactly the same problem appears in another guise, that of the solution of linear equations. The linear transformation T can be written as m linear expressions in $x_1, x_2, \dots x_n$ equated respectively to $y_1, y_2, \dots y_m$. Assign constant values $b_1, b_2, \dots b_m$ to the y's. Then we obtain m linear equations in the x's. Can we find the x's, i.e. solve the linear equations? We can, if we have already inverted T. For then the x's are given in terms of the y's, and assigning the particular values (the b's) to the y's we have the x's which solve the linear equations. For example, if $n=m=2$, then the problem of inverting the linear transformation:

$$y_1=a_{11}x_1+a_{12}x_2 \quad \text{and} \quad y_2=a_{21}x_1+a_{22}x_2,$$

is the same as that of solving the linear equations:

$$a_{11}x_1+a_{12}x_2=b_1 \quad \text{and} \quad a_{21}x_1+a_{22}x_2=b_2.$$

Hence, we have the parallel problems of *inverting a linear transformation* and of *solving a set of linear equations*, both of them exercises in linear algebra.

Linear algebra deals with a great variety of other problems. For example, it provides the conditions under which the quadratic form:

$$a_{11}x_1^2 + a_{22}x_2^2 + \ldots + 2a_{12}x_1x_2 + \ldots > 0 \quad \text{for all } x_1, x_2, \ldots .$$

In its turn, this has applications in the calculus, in the problem of determining the maximum or minimum values of a function of several variables, with or without side relations and constraints.

13.2. The structure of vector spaces. The general vector space $V = \{u, v, w, \ldots\}$ over the field $F = \{a, b, c, \ldots\}$ is a set of double composition with the following properties (as in 8.3). The operation of addition is defined within V, the set of vectors being an additive group with all the operational rules (including the commutative one) being valid. The outside set F provides scalars so that a vector v can be multiplied by a scalar a to provide another vector av. Scalar products satisfy an associative rule: $a(bv) = (ab)v$ and two distributive rules: $a(u+v) = au + av$ and $(a+b)v = av + bv$. Particular scalars operate: $1 \times v = v$, $(-1) \times v = -v$ and $0 \times v = 0$. The last means that the scalar zero times any vector gives the vector zero; the use of two zero elements (scalar and vector), with the same notation 0, need cause no trouble.

The algebraic structure, rather than the geometric interpretation, of a vector space is now examined, with emphasis on the linear aspects. One vector v is a *linear combination* of, or depends on, a set $\{v_1, v_2, \ldots\}$ of vectors if there are scalars such that $v = a_1v_1 + a_2v_2 + \ldots$.* From another aspect of the same property, a set $\{v_1, v_2, \ldots\}$ of vectors is *linearly dependent* if scalars, not all zero, can be found so that $a_1v_1 + a_2v_2 + \ldots = 0$, i.e. if some linear combination of the vectors produces the zero vector. Suppose $a_1 \neq 0$, so that the condition for linear dependence can be written:

$$v_1 = \left(-\frac{a_2}{a_1}\right)v_2 + \left(-\frac{a_3}{a_1}\right)v_3 + \ldots = \lambda_2v_2 + \lambda_3v_3 + \ldots \text{ for some scalars } \lambda_2, \lambda_3, \ldots$$

Hence, in a linearly dependent set, at least one vector is a linear combination of the others. On the other hand, a set $\{v_1, v_2, \ldots\}$ of vectors is *linearly independent* if no scalars can be found so that

* Some of the scalars $a_1, a_2, \ldots$ can be zero and v is still said to depend on the full set of vectors $\{v_1, v_2, \ldots\}$, though in this case v also depends on some set of fewer vectors.

$a_1v_1+a_2v_2+\ldots=0$, the trivial case $a_1=a_2=\ldots=0$ being excluded. In such a set, no one vector is a linear combination of the others.

It is not implied that a vector space V has any linearly dependent, or any linearly independent, vectors. This is still to be explored. It is intuitively clear that V may include a few vectors which are linearly independent but that, as more vectors are taken, the risk of linear dependence increases. We look naturally for the *largest* set of independent vectors. On the other hand, there may be sets of vectors in V on which all vectors (and not just one) of V depend, i.e. a set $\{v_1, v_2, \ldots\}$ such that every vector v is a linear combination of $v_1, v_2, \ldots$: $v=\lambda_1v_1+\lambda_2v_2+\ldots$ for scalars λ_1, λ_2, ... If such a set exists, then it is said to *span* V. The implication is that a spanning set of vectors is enough to describe V, all vectors being some linear combination of them. Here, we look for a rather large set of vectors to span V and, the fewer vectors we take, the greater the risk that they fail to span V. We look naturally for the *smallest* set of vectors which spans V. These matters are treated quite formally in 15.9, where the following basic result is established.

There may be no set of vectors, however large in number, sufficient to span V. The vector space is then said to be of *infinite dimension*. Otherwise, there is a positive integer n which is both the largest number of linearly independent vectors in V and the smallest number of vectors spanning V. More specifically, there exists a set $\{v_1, v_2, \ldots v_n\}$ of n vectors such that the vectors are linearly independent and span V. The vector space is said to be of *dimension* n and such a set of n vectors is said to be a *basis* of the space. The corollary is equally useful in practice: no set of more than n vectors of V can be linearly independent; no set of fewer than n vectors can span V. In particular, given $n+1$ vectors, they must be linearly dependent; given only $n-1$ vectors, they are not enough to span V.

These general ideas can be applied to the special and practical case of a space $V_n(F)$ of n-tuples over a field F (usually real numbers). The field F is used to provide both the scalar a and the components $x_1, x_2, \ldots x_n$ of the n-tuple vector v. The first feature of $V_n(F)$ to emphasise, and the one of most immediate use in practice, is the way in which sums and scalar products are defined and hence the way in which linear combinations of n-tuple vectors are constructed.

Definition: *In the space $V_n(F)$ of n-tuples, the* **sum** *of two n-tuples $v=(x_1, x_2 \dots x_n)$ and $w=(y_1, y_2, \dots y_n)$ is defined as:*

$$v+w=(x_1+y_1, x_2+y_2, \dots x_n+y_n)$$

and the **scalar product** *of a and v is defined as:*

$$av=(ax_1, ax_2, \dots ax_n).$$

This is very simple: to add n-tuple vectors, we add the components separately; to multiply an n-tuple vector by a scalar, we multiply each component by the scalar.

Repeated applications of the specified operations serve to build up a linear combination of n-tuple vectors. Write m vectors:

$$v_1=(x_{11}, x_{21}, \dots x_{n1});\ v_2=(x_{12}, x_{22}, \dots x_{n2});\ \dots\ v_m=(x_{1m}, x_{2m}, \dots x_{nm})$$

and a corresponding set of m scalars: $a_1, a_2, \dots a_m$. Then:

$$v=a_1v_1+a_2v_2+\dots+a_mv_m \quad \dots\dots\dots\dots\dots\dots\dots\dots(1)$$

is a linear combination of the m vectors. The problem is to write the components of $v=(x_1, x_2, \dots x_n)$ separately. By the scalar product rule:

$$\begin{aligned} a_1v_1&=(a_1x_{11}, a_1x_{21}, \dots a_1x_{n1})\\ a_2v_2&=(a_2x_{12}, a_2x_{22}, \dots a_2x_{n2})\\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\\ a_mv_m&=(a_mx_{1m}, a_mx_{2m}, \dots a_mx_{nm}). \end{aligned}$$

By the sum rule, the components of v are simply the sums of the separate components of $a_1v_1, a_2v_2, \dots a_mv_m$, written above in vertical line. Hence:

$$\left.\begin{aligned} x_1&=a_1x_{11}+a_2x_{12}+\dots+a_mx_{1m}\\ x_2&=a_1x_{21}+a_2x_{22}+\dots+a_mx_{2m}\\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\\ x_n&=a_1x_{n1}+a_2x_{n2}+\dots+a_mx_{nm} \end{aligned}\right\} \quad \dots\dots\dots\dots\dots\dots\dots\dots(2)$$

Hence, writing the linear combination (1) for n-tuples simply means that each component separately is the same linear combination as given by (2). This is so basic, and so important in practice, that it is useful to set it out formally:

Theorem: *In the space $V_n(F)$ of n-tuples, sums and scalar products are so defined that a linear combination (1) of n-tuples implies corresponding linear combinations (2) of the components separately.*

Hence, at its lowest, the form (1) is a convenient short-hand for the more detailed (2). Whenever we write a linear combination of n-tuples (1), we can always 'equate components' separately as in (2). Conversely, if we are given a set of equations of form (2), we can always shorten it into n-tuple vector form (1). There is, however, more in the analysis than this; once appropriate algebraic rules are developed, we find it much easier to operate with (1) than with (2).

One application of (1) and (2) is immediate. If $v_1, v_2, \dots v_m$ are linearly dependent, then scalars $a_1, a_2, \dots a_m$ (not all zero) exist so that (1) is the zero vector $0 = (0, 0, \dots 0)$. This means that all the n expressions on the right of (2) are zero. On the other hand, if the vectors are linearly independent, then no such scalars exist. The n expressions of (2) cannot be all zero together, i.e. for any scalars $a_1, a_2, \dots a_m$ (not all zero) at least one of the expressions is non-zero.

The dimension of the space $V_n(F)$ of n-tuples, and a convenient basis for the space, are easily found. The three examples of 13.1 illustrate:

(i) The vector space of real numbers over the field of rationals has plenty of linear combinations, i.e. real numbers such as

$$\surd 2(\surd 2 + 1) = 2 \times 1 + 1 \times \surd 2,$$

a linear combination of 1 and $\surd 2$. There are also plenty of linearly independent sets of real numbers, e.g. 1, $\surd 2$ and $\surd 3$ which are such that no rational a, b and c exist for $a + b\surd 2 + c\surd 3 = 0$. But no finite set of real numbers $x_1, x_2, \dots x_n$ can be found so that every real number x is a (rational) combination of them:

$$x = a_1x_1 + a_2x_2 + \dots + a_nx_n.$$

This is because of the inadequacy of *rational* multiples to take care of the multiplicity of *irrationals*, including surds like $\surd 2$ or $\surd 3$ and transcendentals like π or e. The space is of infinite dimension; it is not a case of $V_n(F)$.

(ii) The vector space of real number pairs (x, y), or of complex numbers $z = x + iy$, over the field of real numbers, is an instance of $V_2(F)$. By the rules for addition and scalar multiplication, any number pair can be expressed:

$$\left.\begin{array}{l}(x, y) = x(1, 0) + y(0, 1) = x\epsilon_1 + y\epsilon_2 \\ \text{where } \epsilon_1 = (1, 0) \quad \text{and} \quad \epsilon_2 = (0, 1)\end{array}\right\} \quad \dots\dots\dots\dots\dots(3)$$

Hence the two vectors $\{\epsilon_1, \epsilon_2\}$ span the space. Moreover, they are linearly independent since $a_1\epsilon_1+a_2\epsilon_2=a_1\,(1, 0)+a_2(1, 0)=(a_1, a_2)\neq 0$ (except in the excluded case $a_1=a_2=0$). The vector space is of dimension 2 and a basis is provided by ϵ_1 and ϵ_2. (3) shows the dependence of every vector (x, y) on the basis. This is not the only basis; indeed almost any pair of vectors can serve as a basis, e.g. $\eta_1=(1, 1)$ and $\eta_2=(1, -1)$ is a basis with (3) replaced by:

$$(x, y)=\tfrac{1}{2}(x+y)\eta_1+\tfrac{1}{2}(x-y)\eta_2.$$

Further insight is obtained by representing the number pairs (x, y) in graphical terms, i.e. as points referred to axes Oxy (Fig. 13.2). Let P_1 and P_2 be two given vectors. They are linearly independent as long as they do not lie on the same radius through O. For, linear dependence of P_1 and P_2 implies P_2 is a multiple of P_1, i.e. OP_1 and OP_2 are the same radius. Moreover, any other point P can be expressed as a linear combination of P_1 and P_2, by multiplying by scalars to turn P_1 into Q_1 and P_2 into Q_2 in such a way that OP is that resultant (i.e. the vector sum) of OQ_1 and OQ_2. Hence

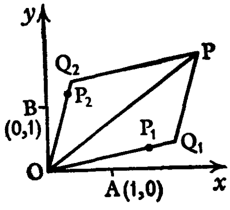


FIG. 13.2

$$OP=\lambda_1 OP_1+\lambda_2 OP_2,$$

for $\lambda_1=OQ_1/OP_1$ and $\lambda_2=OQ_2/OP_2$. Hence, as long as OP_1 and OP_2 are distinct directions, the vectors P_1 and P_2 serve as a basis for the vector space; this is what is meant by 'almost any pair' specified above. The particular basis suggested for use is the pair of points A (1, 0) on Ox and B (0, 1) on Oy. These mark off the units of measurement on the axes and any point P (x, y) is expressed as $x(1, 0)+y(0, 1)$, i.e. x units along Ox and y units along Oy.

(iii) The vector space of quadratics ax^2+bx+c, with real coefficients, over the field of real numbers, corresponds to real number triples (a, b, c) and it is a case of $V_3(F)$. The corresponding relation to (3) is here:

$$\left.\begin{aligned}(a, b, c)&=a(1, 0, 0)+b(0, 1, 0)+c(0, 0, 1)\\ &=a\epsilon_1+b\epsilon_2+c\epsilon_3\end{aligned}\right\}\quad\ldots\ldots\ldots\ldots(4)$$

so that the space is of dimension 3 and the three tuples $\epsilon_1=(1, 0, 0)$, $\epsilon_2=(0, 1, 0)$ and $\epsilon_3=(0, 0, 1)$ provide a basis for the space. Explicitly in terms of quadratics, the basis is $\epsilon_1=x^2$, $\epsilon_2=x$ and $\epsilon_3=1$. The dependence of any quadratic on the basis is given by (4), translated into:

$$ax^2+bx+c=a\times x^2+b\times x+c\times 1.$$

There is a three-dimensional graphical representation, similar to Fig. 13.2 but with the extra dimension. A quadratic or triple (a, b, c) is shown by a point P (a, b, c) referred to axes $Oabc$. Three points P_1, P_2 and P_3 are linearly dependent if they all lie in one plane through O; they are linearly independent (and available as a basis) if they do not. The basis suggested is given by the three points A (1, 0, 0), B (0, 1, 0) and C (0, 0, 1) which are at unit distances along the three axes.

The general result suggested by these examples, and established formally in 15.9, is the following:

THEOREM: *The space $V_n(F)$ of n-tuple vectors $(x_1, x_2, \dots x_n)$ over the field F has dimension n and any set of n vectors, which are both linearly independent and span $V_n(F)$, can be used as a basis. A set of more than n vectors cannot be linearly independent. A set of fewer than n vectors cannot span $V_n(F)$.*

A convenient basis for $V_n(F)$ is provided by the n vectors:*

$$\epsilon_1=(1, 0, 0, \dots 0),\ \epsilon_2=(0, 1, 0, \dots 0),\ \dots\epsilon_n=(0, 0, 0, \dots 1)$$

and the dependence of any vector on the basis is given by:

$$(x_1, x_2, \dots x_n)=x_1\epsilon_1+x_2\epsilon_2+\dots+x_n\epsilon_n \quad \dots\dots\dots\dots\dots(5)$$

Here, (5) for $V_n(F)$ is an obvious extension of (3) for $V_2(F)$ and (4) for $V_3(F)$. The theorem shows that, while $\epsilon_1, \epsilon_2, \dots \epsilon_n$ are linearly independent, the addition of one more vector produces a set which is not linearly independent. Moreover, while $\epsilon_1, \epsilon_2, \dots \epsilon_n$ do span $V_n(F)$, any smaller set of vectors cannot do so.

13.3. Linear transformations and linear equations.

A transformation is a mapping and a linear transformation is that particular case which maps one vector space into another vector space, carrying over a linear combination of vectors from one space to the other. The linear transformation usually taken is a mapping of the space

* The space $V_n(F)$ has zero vector (0, 0, ... 0) for the operation $+$, but there is no unity since no operation $\times$ is implied. However, the n vectors $\epsilon_1, \epsilon_2, \dots \epsilon_n$ can be called *unit vectors* in $V_n(F)$.

$V_n(F)$ of n-tuples into the space $V_m(F)$ of m-tuples. This is examined formally in 15.9 and here in particular cases.

As a simple basic case, though not quite the simplest, take a linear transformation T which maps $V_3(F)$ into itself, i.e. which is a mapping of three dimensions into three dimensions. The transformation is from a triple (x_1, x_2, x_3) of $V_3(F)$ into another triple (y_1, y_2, y_3) of $V_3(F)$. The problem is to specify T and to show how the y's are obtained from the x's. The essential feature of T is that a linear combination of triples of x's must be mapped into the same linear combination of triples of y's.

The specification of T is made as follows. As the basis for $V_3(F)$ write the three vectors $\epsilon_1=(1, 0, 0)$, $\epsilon_2=(0, 1, 0)$ and $\epsilon_3=(0, 0, 1)$. Then:

$$(x_1, x_2, x_3)=x_1\epsilon_1+x_2\epsilon_2+x_3\epsilon_3 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

gives the dependence of any vector of $V_3(F)$ on the basis. Now ϵ_1 is a particular vector which is sent by T into some other particular vector. Let the image of ϵ_1 under T be the vector (a_{11}, a_{21}, a_{31}), where T gives the scalar a's. Similarly, let ϵ_2 have image (a_{12}, a_{22}, a_{32}) and ϵ_3 have image (a_{13}, a_{23}, a_{33}). Altogether there are $3\times 3=9$ scalars specified by T, the constants a_{rs} for r and $s=1, 2, 3$. These are, in fact, a *complete* specification of the linear transformation T and the image (y_1, y_2, y_3) of any vector (x_1, x_2, x_3) can be written in terms of the 9 a's. We are given, therefore, under T:

$$\epsilon_1\rightarrow(a_{11}, a_{21}, a_{31});\ \epsilon_2\rightarrow(a_{12}, a_{22}, a_{32});\ \epsilon_3\rightarrow(a_{13}, a_{23}, a_{33}) \quad \ldots\ldots\ldots(2)$$

By the preservation of linear combinations under T and by (1) and (2):

$$\begin{aligned}(x_1, x_2, x_3)=x_1\epsilon_1+x_2\epsilon_2+x_3\epsilon_3\rightarrow x_1(a_{11}, a_{21}, a_{31})&+x_2(a_{12}, a_{22}, a_{32})\\ &+x_3(a_{13}, a_{23}, a_{33}).\end{aligned}$$

The linear combination of vectors on the right is the vector (y_1, y_2, y_3). By the rules for sums and scalar products, as expressed in (1) and (2) of 13.2, we can write:

$$\left.\begin{aligned}y_1&=a_{11}x_1+a_{12}x_2+a_{13}x_3\\ y_2&=a_{21}x_1+a_{22}x_2+a_{23}x_3\\ y_3&=a_{31}x_1+a_{32}x_2+a_{33}x_3\end{aligned}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

for the separate components. Hence, if (x_1, x_2, x_3) is sent into (y_1, y_2, y_3) by T, then the y's are related to the x's by (3). The linear relations (3) are a complete description of the linear transformation T of $V_3(F)$ into itself, the a's being scalars fixed by (2).

We have achieved in (3) an extension of the even simpler case of a linear transformation in two dimensions, from $V_2(F)$ into itself:

$$y_1 = a_{11}x_1 + a_{12}x_2 \quad \text{and} \quad y_2 = a_{21}x_1 + a_{22}x_2$$

as examined in 7.5. The extension is an increase from two to three dimensions on both sides, the x's and the y's alike. This is not necessary; a linear transformation can be from one space $V_n(F)$ of dimension n to another space $V_m(F)$ of different dimension m. Two other simple cases illustrate.

Suppose the linear transformation T maps the triples (x_1, x_2, x_3) of $V_3(F)$ into the pairs (y_1, y_2) of $V_2(F)$. Keeping ϵ_1, ϵ_2 and ϵ_3 as the basis for $V_3(F)$, we specify their images under T:

$$\epsilon_1 \to (a_{11}, a_{21});\ \epsilon_2 \to (a_{12}, a_{22});\ \epsilon_3 \to (a_{13}, a_{23})$$

so that: $$\begin{aligned}(x_1, x_2, x_3) &= x_1\epsilon_1 + x_2\epsilon_2 + x_3\epsilon_3\\ &\to x_1(a_{11}, a_{21}) + x_2(a_{12}, a_{22}) + x_3(a_{13}, a_{23})\\ &= (y_1, y_2)\end{aligned}$$

where $$\left.\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 + a_{13}x_3\\ y_2 &= a_{21}x_1 + a_{22}x_2 + a_{23}x_3\end{aligned}\right\} \quad \text{......................(4)}$$

Hence T is given by (4), completely described by $2 \times 3 = 6$ scalars, the a's.

On the other hand, suppose T maps the pairs (x_1, x_2) of $V_2(F)$ into the triples (y_1, y_2, y_3) of $V_3(F)$. The basis for $V_2(F)$ is the pair of vectors $\epsilon_1 = (1, 0)$ and $\epsilon_2 = (0, 1)$ with images under T taken as:

$$\epsilon_1 \to (a_{11}, a_{21}, a_{31}) \quad \text{and} \quad \epsilon_2 \to (a_{12}, a_{22}, a_{32})$$

and: $$\begin{aligned}(x_1, x_2) &= x_1\epsilon_1 + x_2\epsilon_2 \to x_1(a_{11}, a_{21}, a_{31}) + x_2(a_{12}, a_{22}, a_{32})\\ &= (y_1, y_2, y_3)\end{aligned}$$

where $$\left.\begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2\\ y_2 &= a_{21}x_1 + a_{22}x_2\\ y_3 &= a_{31}x_1 + a_{32}x_2\end{aligned}\right\} \quad \text{........................(5)}$$

T given by (5) is described by $3 \times 2 = 6$ scalars, the a's.

The generalisation is now clear enough. If the linear transformation T maps $V_n(F)$ into $V_m(F)$, write the images of the basis for $V_n(F)$:

$$\epsilon_1 \to (a_{11}, a_{21}, \ldots a_{m1});\ \epsilon_2 \to (a_{12}, a_{22}, \ldots a_{m2});\ \ldots\ \epsilon_n \to (a_{1n}, a_{2n}, \ldots a_{mn})$$

and, if $(x_1, x_2, \ldots x_n) \to (y_1, y_2, \ldots y_m)$, then:

$$\left.\begin{array}{l} y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ y_m = a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \end{array}\right\} \ldots\ldots\ldots\ldots\ldots\ldots(6)$$

Hence, the general form (6) of the linear transformation, from $V_n(F)$ to $V_m(F)$, is specified by $m \times n$ scalars, the a's, as given by the images of the basis $\epsilon_1, \epsilon_2, \ldots \epsilon_n$ taken for $V_n(F)$. The particular cases, (3), (4) and (5), have $m=n$, $m<n$ and $m>n$ respectively.

In the linear transformation (6), suppose that we know that $(x_1, x_2, \ldots x_n)$ of $V_n(F)$ maps into a particular and specific vector $(b_1, b_2, \ldots b_m)$ of $V_m(F)$. Then the n-tuple of x's is such that:

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n = b_m \end{array}\right\} \ldots\ldots\ldots\ldots\ldots\ldots(7)$$

The problem is to find the x's, i.e. the solution of the set of linear equations (7). Linear equations are a particular aspect of linear transformations.

Hence, we have a dual problem: to invert (6) and to obtain the x's in terms of the y's; or to solve (7) and to obtain the x's which satisfy the equations for given b's. In both cases, the set of $m \times n$ scalars, the a's, is given, the structure of the transformation or equation system. If one problem is solved, so is the dual.

As a preliminary canter, we can examine the solution of the problem in the simple case $m=n=2$. As shown in 7.5, elimination of one variable to get the second is the method to follow. So the inverse of the linear transformation, $y_1 = a_{11}x_1 + a_{12}x_2$ and $y_2 = a_{21}x_1 + a_{22}x_2$, is:

$$\left.\begin{array}{c} \dfrac{x_1}{a_{22}y_1 - a_{12}y_2} = \dfrac{x_2}{a_{11}y_2 - a_{21}y_1} = \dfrac{1}{a_{11}a_{22} - a_{12}a_{21}} \\ \text{where } A = a_{11}a_{22} - a_{12}a_{21} \neq 0 \end{array}\right\} \ldots\ldots\ldots\ldots\ldots(8)$$

Equally, the solution of the equations:

$$a_{11}x_1+a_{12}x_2=b_1 \quad\text{and}\quad a_{21}x_1+a_{22}x_2=b_2$$

is:

$$\left.\begin{array}{c}\dfrac{x_1}{a_{22}b_1-a_{12}b_2}=\dfrac{x_2}{a_{11}b_2-a_{21}b_1}=\dfrac{1}{a_{11}a_{22}-a_{12}a_{21}}\\ \text{where } A=a_{11}a_{22}-a_{12}a_{21}\neq 0\end{array}\right\}\dots\dots\dots\dots(9)$$

Here, (8) becomes (9) simply by assigning $y_1=b_1$ and $y_2=b_2$. In (8), x_1 and x_2 are linear expressions in the y's, the inverse of the linear transformation:

$$x_1=\frac{1}{A}\left(a_{22}y_1-a_{12}y_2\right) \quad\text{and}\quad x_2=\frac{1}{A}\left(-a_{21}y_1+a_{11}y_2\right).$$

In (9), x_1 and x_2 take specified values, the solution of the equations. In each interpretation, the solution of the problem is given in the *main case* where $A\neq 0$. There remains the *degenerate case* where $A=0$, and where the linear transformation cannot be inverted, or the equations solved, at least completely (see 13.9 Ex. 29).

The algebraic method used, i.e. the successive elimination of variables until only one is left, can be extended, though with increasing labour, to cases of more than two variables. In 13.9 Ex. 4, the inversion of (3) in the 3 × 3 case is achieved. In 13.9 Ex. 5, a similar attack is made on the 2 × 3 case of (4) and on the 3 × 2 case of (5). A difficulty arises; there is now a 'surplus' variable, one variable too many, e.g. x_3 in (4) and y_3 in (5). A solution of the problem is only attained if the surplus variable is given an assigned value.

This suggests two things. First, something must be done to simplify the algebra in inverting the general linear transformation (6), or in solving the general system of equations (7), to avoid having to slog out the result in each particular case. The matrix notation is introduced for these (and many other) purposes. Second, care must be taken to distinguish the various possibilities and, in particular, to keep a sharp eye open for any degenerate cases.

13.4. Matrices. A matrix is a notation for an ordered set of $m\times n$ elements, arranged in an array of m rows and n columns, for given positive integers m and n. The elements can be entities of any kind, typically real scalars, but also including such entities as polynomials or functions (13.9 Ex. 6). In the following development, however, it is taken for convenience that the elements are from some field F of scalars.

Definition: *A* **matrix** *of order $m \times n$ is a set of elements in m rows and n columns:*

$$\mathbf{A} = \| a_{rs} \| = \left\| \begin{array}{llll} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ . & . & . & . \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{array} \right\|$$

where $r = 1, 2, \dots m$ denote the rows and $s = 1, 2, \dots n$ the columns.

There are two alternative notations. In one, a single letter is used for a matrix, printed in bold type $\mathbf{A}$ to indicate that it is a complex of values (an array of m rows and n columns) and not a single value. In the other, the elements of the matrix are spelled out in their $m \times n$ array and bordered by double vertical lines. Variants of this second notation are used:

$$\left\| \begin{array}{llll} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ . & . & . & . \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{array} \right\| \qquad \left(\begin{array}{llll} a_{11} & a_{12} & \dots & a_{1n} \\ a_{22} & a_{22} & \dots & a_{2n} \\ . & . & . & . \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{array} \right) \qquad \left[\begin{array}{llll} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ . & . & . & . \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{array} \right]$$

all indicate the same matrix. In the shortened form of this notation, i.e. $\| a_{rs} \|$, (a_{rs}) or $[a_{rs}]$, the general subscripts need to be specified: $r = 1, 2, \dots m$ and $s = 1, 2, \dots n$.

Two special cases of matrices arise when $m = 1$ and when $n = 1$. In the first case, we obtain a *row vector* which is an n-tuple of $V_n(F)$ where F is the field from which the elements are drawn:

Matrix, order $1 \times n$ = row vector $\| a_1 \, a_2 \dots a_n \| = (a_1 \, a_2 \dots a_n)$.

In the second case, we have a *column vector*, an m-tuple of $V_m(F)$:

Matrix, order $m \times 1$ = column vector $\left\| \begin{array}{c} a_1 \\ a_2 \\ \dots \\ a_m \end{array} \right\| = \{a_1 \, a_2 \dots a_m\}$.

The alternative notation, to the $\| \dots \|$ matrix notation, is to write

(...) for a row vector and {...} for a column vector; this saves space while distinguishing row from column vectors.

A matrix of order $m \times n$ is built up from vectors, i.e. m row vectors each of order $1 \times n$, or n column vectors each of order $m \times 1$. It also gives rise to various other vectors. For example, the matrix $\| a_{rs} \|$ of order $n \times n$ has for its *leading diagonal* the n-tuple vector $(a_{11}, a_{22}, \dots a_{nn})$.

Two particular matrices of a given order require separate notations:

NOTATION: *The* **zero matrix** *of order $m \times n$ is $\mathbf{O}_{mn}$, consisting of m rows and n columns of elements, all zero. The* **unit matrix** *of order $n \times n$ is* $\mathbf{I}_n = \left\| \begin{matrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ . & . & . & . \\ 0 & 0 & \dots & 1 \end{matrix} \right\|$ *comprising a leading diagonal of elements all unity, and zero elements elsewhere.*

If the order is understood, without ambiguity, the zero matrix can be denoted **O** and the unit matrix **I**.

When the elements of matrices are summed in various ways, the Σ notation of 1.7 is of particular use. For example, the sum of the elements of the rth row of $\mathbf{A} = \| a_{rs} \|$ can be written $\sum_{s=1}^{n} a_{rs}$ for $r = 1, 2, \dots m$. The most important use of the Σ notation here is in writing *inner products*; these are scalar values which appear in multiplying matrices. The notation can be given for vectors as well as for matrices generally:

NOTATION: *Two vectors $a = (a_1, a_2, \dots a_n)$ and $b = (b_1, b_2, \dots b_n)$, each of n elements, give the* **inner product**:

$$a \,.\, b = \sum_{s=1}^{n} a_s b_s = a_1 b_1 + a_2 b_2 + \dots + a_n b_n.$$

Two matrices $\mathbf{A} = \| a_{rt} \|$ of order $m \times k$ and $\mathbf{B} = \| b_{ts} \|$ of order $k \times n$ are **conformable** *and give $m \times n$* **inner products**:

$$\sum_{t=1}^{k} a_{rt} b_{ts} = a_{r1} b_{1s} + a_{r2} b_{2s} + \dots + a_{rk} b_{ks}$$

for $r = 1, 2, \dots m$ and $s = 1, 2, \dots n$.

The definition of conformable matrices, implicit in this notation, is to be noticed. Not all pairs of matrices are conformable by any means. It is necessary that **A** has the same number (k) of columns as **B** has rows if **A** is to be conformable with **B**. Moreover, matrices are conformable in a particular order, i.e. **A** of order $m \times k$ with **B** of order $k \times n$. Reverse the order and the matrices are not generally conformable, i.e. **B** of order $k \times n$ is not conformable with **A** of order $m \times k$ (except in case where $m = n$). It is always essential, before writing inner products, to check that the matrices used are conformable and to stick to the order in which they are conformable.

13.5. Operational rules for matrices. Notations are always subject to specified algebraic operations, usually of a simple kind. The matrix notation is quite exceptional in that it is handled algebraically by a very elaborate system of operations, the subject of matrix algebra. A matrix is a new symbol, for an entity of a new kind and one of very wide scope. We are at liberty to define operations on matrices in any way we find convenient. Amongst those we choose to define are operations labelled 'addition' and 'multiplication' of matrices. They are specified with an eye on the applications of matrices, particularly to linear transformations and linear equations. However, as in Boolean algebra of Chapter 4, choice of the labels is made as a way out of a dilemma. The operation of 'multiplication' of matrices, in particular, does not satisfy all the familiar multiplicative rules, as the label might suggest. It is *not* the same kind of operation as the multiplication of real or complex numbers, or of the elements of a field generally. Not only is the commutative rule invalid, but matrix products generally lack reciprocals and fail to meet the cancellation rule. It might be preferable to use a different term, e.g. to talk of the 'conformation' of matrices, rather than their 'multiplication', since the operation is limited to matrices which are conformable in the sense of 13.4. However, 'multiplication' is the label in general use, and we must employ it, though with great care in remembering that not all the usual multiplicative rules apply.

Four sets of definitions are required for matrices of various orders:

DEFINITION: **Equality and inequality.** *If* $\mathbf{A} = \| a_{rs} \|$ *and* $\mathbf{B} = \| b_{rs} \|$ *are of the same order* $m \times n$, *then* $\mathbf{A} = \mathbf{B}$ *if* $a_{rs} = b_{rs}$ *and* $\mathbf{A} > \mathbf{B}$ *if* $a_{rs} > b_{rs}$ *for all* $r = 1, 2, \ldots m$ *and* $s = 1, 2, \ldots n$.

As a particular case, take $\mathbf{B}=\mathbf{0}$ of order $m \times n$. Then $\mathbf{A}=\mathbf{0}$ means that each element of $\mathbf{A}$ is zero and $\mathbf{A}>\mathbf{0}$ that each element of $\mathbf{A}$ is positive. As opposed to the positive matrix $\mathbf{A}>\mathbf{0}$, we can write the non-negative matrix $\mathbf{A}\geqslant\mathbf{0}$, meaning $a_{rs}\geqslant 0$ for all r and s (except that the case *all* $a_{rs}=0$ is excluded). Hence $\mathbf{A}\geqslant\mathbf{0}$ covers a range of possibilities: the case where $a_{rs}>0$ for all r and s, together with cases where $a_{rs}>0$ for some r and s and $a_{rs}=0$ for other r and s. (See 13.9, Ex. 10.)

DEFINITION: **Addition and scalar products.** *If* $\mathbf{A}=\|\,a_{rs}\,\|$ *and* $\mathbf{B}=\|\,b_{rs}\,\|$ *are of the same order* $m \times n$, *then the* **sum** $\mathbf{A}+\mathbf{B}$ *is the matrix* $\mathbf{C}=\|\,c_{rs}\,\|$ *where* $c_{rs}=a_{rs}+b_{rs}$, *and the* **scalar product** $\lambda\mathbf{A}$ *is the matrix* $\mathbf{D}=\|\,d_{rs}\,\|$ *where* $d_{rs}=\lambda a_{rs}$, *for all* $r=1, 2, \ldots m$ *and* $s=1, 2, \ldots n$.

These definitions are simple enough. Any matrix can be multiplied by a scalar λ; it is simply a matter of multiplying each element by λ. Any two matrices of the same order can be added; each element of one matrix is simply added to the corresponding element of the other. Within the set of all matrices of a given order $m \times n$, the operations of addition and scalar multiplication are closed. On the other hand, there is no meaning attached to the addition of matrices of different order.

A direct consequence of the definition is that all the rules for the operations of sums and scalar products are satisfied:

THEOREM: *The set of all matrices of a given order* $m \times n$ *is a vector space over the field from which scalars are drawn.*

To prove, it is a matter of checking the rules from the definition of sums and scalar products. The set of all $m \times n$ matrices is closed under sums and scalar products, sums are commutative and associative:

$$\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A} \quad \text{and} \quad \mathbf{A}+(\mathbf{B}+\mathbf{C})=(\mathbf{A}+\mathbf{B})+\mathbf{C}$$

and scalar products are associative and distributive:

$$\lambda(\mu\mathbf{A})=(\lambda\mu)\mathbf{A};\ \lambda(\mathbf{A}+\mathbf{B})=\lambda\mathbf{A}+\lambda\mathbf{B};\ (\lambda+\mu)\mathbf{A}=\lambda\mathbf{A}+\mu\mathbf{A}.$$

There is an identity element for addition, the zero vector $\mathbf{0}$ of order $m \times n$:

$$\mathbf{A}+\mathbf{0}=\mathbf{0}+\mathbf{A}=\mathbf{A}$$

and there is an additive inverse, the negative $-\mathbf{A}$ of $\mathbf{A}$:

$$\mathbf{A}+(-\mathbf{A})=(-\mathbf{A})+\mathbf{A}=\mathbf{0}.$$

Here, if $\mathbf{A}=\|a_{rs}\|$, then $-\mathbf{A}=\|(-a_{rs})\|$ and the same negative $-\mathbf{A}$ is obtained by multiplication of $\mathbf{A}$ by the scalar -1. With a negative defined, the difference of two vectors follows at once: $\mathbf{A}-\mathbf{B}=\mathbf{A}+(-\mathbf{B})$. If $\mathbf{A}=\|a_{rs}\|$ and $\mathbf{B}=\|b_{rs}\|$, then

$$\mathbf{A}-\mathbf{B}=\|(a_{rs}-b_{rs})\|.$$

Finally, a zero difference is to be associated with equality: $\mathbf{A}-\mathbf{B}=\mathbf{0}$ implies $\mathbf{A}=\mathbf{B}$.

DEFINITION: **Multiplication.** *If* $\mathbf{A}=\|a_{rt}\|$ *of order* $m\times k$ *is conformable with* $\mathbf{B}=\|b_{ts}\|$ *of order* $k\times n$, *then the* **product AB** *is the matrix* $\mathbf{C}=\|c_{rs}\|$ *of order* $m\times n$ *where* c_{rs} *is the inner product:*

$$c_{rs}=\sum_{t=1}^{k}a_{rt}b_{ts} \quad \textit{for } r=1,\,2,\,\ldots\, m \textit{ and } s=1,\,2,\,\ldots\, n.$$

This definition appears complicated, but it is deliberately designed to agree with successive applications of linear transformations. For example:

Write two matrices

$$\mathbf{A}=\left\|\begin{matrix}a_{11} & a_{12}\\ a_{21} & a_{22}\end{matrix}\right\| \quad \text{and} \quad \mathbf{B}=\left\|\begin{matrix}b_{11} & b_{12}\\ b_{21} & b_{22}\end{matrix}\right\|$$

of order 2×2. They are conformable and yield, by the definition, the product

$$\mathbf{AB}=\left\|\begin{matrix}c_{11} & c_{12}\\ c_{21} & c_{22}\end{matrix}\right\|$$

where:

$$c_{11}=a_{11}b_{11}+a_{12}b_{21} \qquad c_{12}=a_{11}b_{12}+a_{12}b_{22}$$
$$c_{21}=a_{21}b_{11}+a_{22}b_{21} \qquad c_{22}=a_{21}b_{12}+a_{22}b_{22}$$

Now take the linear transformations:

$$z_1=a_{11}y_1+a_{12}y_2 \quad \text{and} \quad y_1=b_{11}x_1+b_{12}x_2$$
$$z_2=a_{21}y_1+a_{22}y_2 \qquad\qquad y_2=b_{21}x_1+b_{22}x_2$$

and combine them (in succession) to give:

$$z_1=a_{11}(b_{11}x_1+b_{12}x_2)+a_{12}(b_{21}x_1+b_{22}x_2)$$
$$z_2=a_{21}(b_{11}x_1+b_{12}x_2)+a_{22}(b_{21}x_1+b_{22}x_2)$$

i.e.

$$z_1=c_{11}x_1+c_{12}x_2$$
$$z_2=c_{21}x_1+c_{22}x_2.$$

The two given linear transformations have matrices **A** and **B** of coefficients. The combination (produet) of these has matrix **AB**.

To write the product of matrices **AB** in practice, the first step is check that **A** is conformable with **B**, i.e. order $m \times k$ combines with order $k \times n$ (the k being common) to give order $m \times n$. Then the $m \times n$ elements of **AB** are written down in succession by the rule of inner products, a rule which can be described as 'across and down'. The element c_{rs} of **AB** is got by reading across the rth row of **A** and down the sth column of **B**, multiplying corresponding elements and adding the results. So c_{11} is across the first row of **A** and down the first column of **B**; c_{12} is across the first row of **A** and down the second column of **B**; and so on.

Notice, in particular, that it is only in special cases that both **AB** and **BA** are defined, and that (even if they are) they are not generally the same matrices. Products of matrices are *not* to be taken as commutative. So, if **A** is of order $m \times k$ and **B** of order $k \times n$, then **AB** can be written. But **B** of order $k \times n$ is not conformable with **A** of order $m \times k$ if $m \neq n$, and no product **BA** can be written. If $m = n$, then **BA** exists, but it is not generally the same matrix as **AB**. Some examples illustrate:

(i)
$$\begin{Vmatrix} 1 & -1 \\ 1 & 0 \\ 0 & 1 \end{Vmatrix} \times \begin{Vmatrix} 0 & 1 \\ -1 & 0 \end{Vmatrix} = \begin{Vmatrix} 1 & 1 \\ 0 & 1 \\ -1 & 0 \end{Vmatrix}$$

but the product in reverse order does not exist. Notice that the given matrices are of order 3×2 and 2×2, which are conformable and give a product $\| c_{rs} \|$ of order 3×2. The inner product, or across and down, rule then provides:

$$c_{11} = 1 \times 0 + (-1) \times (-1) = 1; \quad c_{12} = 1 \times 1 + (-1) \times 0 = 1;$$

and so on.

(ii)
$$\begin{Vmatrix} 1 & -1 \\ 1 & 0 \end{Vmatrix} \times \begin{Vmatrix} 0 & 1 \\ -1 & 0 \end{Vmatrix} = \begin{Vmatrix} 1 & 1 \\ 0 & 1 \end{Vmatrix}$$

and
$$\begin{Vmatrix} 0 & 1 \\ -1 & 0 \end{Vmatrix} \times \begin{Vmatrix} 1 & -1 \\ 1 & 0 \end{Vmatrix} = \begin{Vmatrix} 1 & 0 \\ -1 & 1 \end{Vmatrix}.$$

Both products exist, but they are different matrices.

(iii)
$$\begin{Vmatrix} 1 & -1 \\ 1 & 0 \\ 0 & 1 \end{Vmatrix} \times \begin{Vmatrix} 0 \\ -1 \end{Vmatrix} = \begin{Vmatrix} 1 \\ 0 \\ -1 \end{Vmatrix}$$

and
$$\begin{Vmatrix} 1 & 0 & -1 \end{Vmatrix} \times \begin{Vmatrix} 1 & -1 \\ 1 & 0 \\ 0 & 1 \end{Vmatrix} = \begin{Vmatrix} 1 & -2 \end{Vmatrix}$$

are two (different) cases illustrating that a matrix can be multiplied by appropriate row or column vectors.

Since only particular matrices (those which happen to conform) have products, there is generally no question that a comprehensive set of matrices is closed under multiplication, no question that the set is a ring or a field. In particular, in the set of all matrices of given order $m \times n$, no products are defined at all when $m \neq n$. The case where $m = n$ is considered later (13.6).

However, for these matrices **A, B, C**, ... which do happen to conform, the following properties hold (13.9 Ex. 16–19):

Associative: $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$

Distributive: $\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC}$ and $(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}$

Zero matrices: $\mathbf{AO}_{nn} = \mathbf{O}_{mm}\mathbf{A} = \mathbf{O}_{mn}$ (**A** of order $m \times n$)

Unit matrices: $\mathbf{AI}_n = \mathbf{I}_m\mathbf{A} = \mathbf{A}$ (**A** of order $m \times n$).

These are much used in practice. On the other hand, apart from the non-commutative feature of products of matrices, it is generally the case that reciprocals are lacking and that cancellation is not valid. It may be found, for conformable matrices, that $\mathbf{AB} = \mathbf{0}$; this does not imply that either $\mathbf{A} = \mathbf{0}$ or $\mathbf{B} = \mathbf{0}$. For example, the following two non-zero matrices multiply to zero:

$$\begin{Vmatrix} 1 & 1 \\ 0 & 0 \end{Vmatrix} \times \begin{Vmatrix} 1 & 1 \\ -1 & -1 \end{Vmatrix} = \begin{Vmatrix} 0 & 0 \\ 0 & 0 \end{Vmatrix} = \mathbf{0}.$$

So, in a set of 2×2 matrices (for example), there can be divisors of zero.

The last definition is of a different kind, but a very simple one. It relates to the operation of interchanging the rows and columns of a matrix:

DEFINITION: **Transposition.** *If* $\mathbf{A} = \| a_{rs} \|$ *is of order* $m \times n$, *then the* **transpose**

$$\mathbf{A}' = \| a_{sr} \| \text{ is of order } n \times m.$$

Several properties follow from the definition:

$$(\mathbf{A}+\mathbf{B})' = \mathbf{A}'+\mathbf{B}'; \; (\mathbf{AB})' = \mathbf{B}'\mathbf{A}'; \; (\mathbf{A}')' = \mathbf{A}; \; \mathbf{I}_n' = \mathbf{I}_n$$

for matrices of appropriate orders. The last of these properties, that a unit matrix is unchanged when transposed, raises the question of what kind of matrix **A** has the property: $\mathbf{A}'=\mathbf{A}$. It is easily seen that **A** must be of order $n\times n$ and that its elements are symmetrical about the leading diagonal (13.9 Ex. 20).

13.6. Square matrices. A matrix of order $n\times n$, for a positive integer n, is called a *square matrix*; it has the same number of rows as columns.* The set of all square matrices of a given order would appear to have very tidy properties. All the rules for sums and scalar products still apply, so that the set is still a vector space over the field of scalars. Moreover, every matrix of the set is conformable with every other matrix and the product is also a matrix of the set. If **A** and **B** are of order $n\times n$, then **AB** and **BA** both exist, also as matrices of order $n\times n$. Hence the set should be obedient under the operation of multiplication of matrices, except (as expected) that products are not commutative: **AB** and **BA** exist but $\mathbf{AB}\neq\mathbf{BA}$ in general.

In pursuing this matter, we note that there is no difficulty about either zero or unit matrices of square form:

$$\mathbf{AO}=\mathbf{OA}=\mathbf{O} \quad \text{and} \quad \mathbf{AI}=\mathbf{IA}=\mathbf{A}$$

all matrices written being of order $n\times n$. The remaining question, still to be answered, is whether reciprocals exist, i.e. whether there is a square matrix $\mathbf{A}^{-1}$ to match the given square matrix **A**:

$$\mathbf{AA}^{-1}=\mathbf{A}^{-1}\mathbf{A}=\mathbf{I}.$$

To get an answer, we need to go a good way off course in introducing the concept of a *determinant*. A square matrix $\mathbf{A}=\|a_{rs}\|$ is an ordered array of $n\times n$ scalars (e.g. numerical values).† Then the determinant $A=|\mathbf{A}|=|a_{rs}|$ to be defined is a single scalar to be

* A matrix of order $m\times n$, where $m\neq n$, can likewise be called 'rectangular'. These terms have no geometric content; they refer only to the appearance of the array of elements of the matrix.

† This is the usual case but it can be extended to a square matrix of elements of any kind (e.g. polynomials), in which case the determinant is a single entity of the same kind (e.g. a single polynomial). Determinants arise in quite elementary algebra, in connection with the solution of linear equations. The impression is sometimes conveyed that a matrix is a generalised determinant. It is nothing of the sort. A matrix is an *array* of elements; a determinant is an *algebraic expression* in the elements. These are two quite different concepts.

derived from the $n \times n$ scalars of $\mathbf{A}$. The definition is built up on the principle of mathematical induction; a determinant is first specified very simply in the case $n=1$ and then a rule is laid down to express determinants of order n in terms of those of order $n-1$. A notation is required:

NOTATION: *If $\mathbf{A}=\| a_{rs} \|$ is a matrix of order $n \times n$, write $\mathbf{A}_{rs}$ for the matrix of order $(n-1) \times (n-1)$ obtained by deleting from $\mathbf{A}$ the rth row and the sth column, intersecting in the element a_{rs}. If $A=|\mathbf{A}|$ is the determinant of $\mathbf{A}$, then the* **co-factor** *of a_{rs} in $\mathbf{A}$ is the determinant of $\mathbf{A}_{rs}$ with an approriate sign: $A_{rs}=(-1)^{r+s}|\mathbf{A}_{rs}|$.*

The rule for a determinant A of order n is given in terms of the co-factors A_{rs}, i.e. in terms of determinants of order $n-1$.

DEFINITION: *The* **determinant** *$A=|\mathbf{A}|=|a_{rs}|$ of order n is obtained from a square matrix $\mathbf{A}=\| a_{rs} \|$ of order $n \times n$. When $n=1$, $A=\mathbf{A}=a_{11}$, a single element. The rule for a determinant of order n in terms of determinants of order $n-1$ is:*

$$A=\sum_{t=1}^{n} a_{1t}A_{1t} \qquad \text{.............................}(1)$$

where A_{1t} is the co-factor of a_{1t} $(t=1, 2, \ldots n)$.

From the definition, determinants of successive orders are evaluated step by step. The relation (1) is in terms of the elements of the first row of whatever matrix is considered, and of the appropriate co-factors of these elements. So:

$\underline{n=1}$: $\quad \mathbf{A}=a_{11}$

and $\quad A=a_{11}$

$\underline{n=2}$: $\quad \mathbf{A}=\left\| \begin{matrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{matrix} \right\|$

and $\quad A=a_{11}|a_{22}|-a_{12}|a_{21}|=a_{11}a_{22}-a_{12}a_{21}$.

Here the determinant A is the 'cross-product' $a_{11}a_{22}-a_{12}a_{21}$ of the four elements of $\mathbf{A}$. It is the expression already met in inverting a 2×2 linear transformation or in solving two linear equations in two variables (13.3 above).

$\underline{n=3}$: $\quad \mathbf{A}=\left\| \begin{matrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{matrix} \right\|$

and $$A = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

Using the 'cross-products' for the determinants of second order:

$$A = a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31})$$

i.e.

$$A = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31} - a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} \quad \ldots(2)$$

This appears, as in 13.9 Ex. 4, in the inversion of a 3 × 3 linear transformation.

The process continues for $n=4, 5, \ldots$, by repeated use of (1), and the expansion for A in terms of the a's becomes increasingly involved. However, the pattern is already clear in (2) for $n=3$. In general, for A of order n, there are $n!$ terms in the full expansion, each consisting of n elements, one from each row and one from each column of **A**. By the symmetry of this expression for A, it follows that the expansion rule (1) need not be confined to the first row of elements in **A**; it could equally well be given in terms of the rth row:

$$A = \sum_{t=1}^{n} a_{rt}A_{rt} \quad (r = 1, 2, \ldots n).$$

Further (1) need not be given only for rows; it is equally applicable to columns: $A = \sum\limits_{t=1}^{n} a_{ts}A_{ts}$ $(s = 1, 2, \ldots n)$. Again, from the pattern of plus and minus terms in (2), it appears that $A=0$ if one row of elements in **A** is identical (element by element) with another row. For example, put $a_{11}=a_{21}$, $a_{12}=a_{22}$, $\ldots$ and $A=0$. Making this substitution in (1), we obtain: $\sum\limits_{t=1}^{n} a_{2t}A_{1t}=0$. This implies that, if the elements of the second row of **A** are multiplied by the co-factors of a different row (the first) and the products added, then the result is zero. The result is generalised for any pair of rows: $\sum\limits_{t=1}^{n} a_{rt}A_{st}=0$ $(r \neq s)$; and for any pair of columns: $\sum\limits_{t=1}^{n} a_{ts}A_{tr}=0$ $(r \neq s)$. The two sets of results can be assembled in expansion rules for determinants:

THEOREM: *The* **expansion** *of a determinant $A = | a_{rs} |$ of order n in terms of co-factors proceeds either by rows or by columns:*

$$\left.\begin{array}{lll} \textit{By rows:} & \sum_{t=1}^{n} a_{rt}A_{st}=A\,(r=s)\quad \textit{and} & =0\,(r\neq s) \\ \textit{By columns:} & \sum_{t=1}^{n} a_{ts}A_{tr}=A\,(r=s)\quad \textit{and} & =0\,(r\neq s) \end{array}\right\}\dots\dots\dots\dots\dots(3)$$

The expansion rules are essentially simple: if elements and co-factors of the *same* row or column are taken, the result is A; if elements and co-factors of *different* rows or columns, the result is zero.

Notice that the notation for determinants is an exact parallel of that for matrices, single vertical lines $|\dots|$ being used instead of double lines $||\dots||$. So the matrix $\mathbf{A}=||\,a_{rs}\,||$ gives the determinant $A=|\,a_{rs}\,|$. In full:

$$\mathbf{A}=\begin{Vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ . & . & . & . \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{Vmatrix} \quad \text{and} \quad A=\begin{vmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ . & . & . & . \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{vmatrix}$$

represent a square matrix of order $n \times n$ and its determinant of order n.

Square matrices of order $n \times n$ can be divided into two classes, according as the determinant is zero or non-zero:

DEFINITION: *The square matrix* **A** *with determinant A is* **singular** *if $A=0$ and* **non-singular** *if $A\neq 0$.*

The existence of an inverse matrix $\mathbf{A}^{-1}$ to a given square matrix **A** turns on whether **A** is singular or not. The following definition and theorem establish the position.

DEFINITION: *The square matrix* **A** *of order $n \times n$ has determinant A. If* **A** *is singular ($A=0$), no inverse $\mathbf{A}^{-1}$ exists. If* **A** *is non-singular ($A\neq 0$), then*

$$\mathbf{A}^{-1}=\frac{1}{A}\begin{Vmatrix} A_{sr} \end{Vmatrix}=\frac{1}{A}\begin{Vmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ A_{12} & A_{22} & \dots & A_{n2} \\ . & . & . & . \\ A_{1n} & A_{2n} & \dots & A_{nn} \end{Vmatrix}.$$

where $A_{rs}=(-1)^{r+s}\,|\,\mathbf{A}_{rs}\,|$ is the co-factor of a_{rs} in A.

Hence $\mathbf{A}^{-1}$ is also a matrix of order $n \times n$. By the definition, it is obtained by writing the matrix $||\,A_{rs}\,||$ of the $n \times n$ co-factors, by transposing to $\| A_{sr} \| = \| A_{rs} \|'$ and by dividing every element by A. This requires $A \neq 0$. To identify $\mathbf{A}^{-1}$ as the inverse of $\mathbf{A}$ in the sense of a group under $\times$:

THEOREM: *If* $\mathbf{A}$ *is non-singular, then* $\mathbf{A}^{-1}$ *is the only matrix with the property:*

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}.$$

Proof: $$\mathbf{AA}^{-1} = \| a_{rs} \| \times \| A_{sr} \| / A = \| c_{rs} \|$$

where $$c_{rs} = \frac{1}{A} \sum_{t=1}^{n} a_{rt} A_{st}$$

by the multiplication rule for matrices. But $c_{rs} = \frac{1}{A} A = 1$ $(r = s)$ and $c_{rs} = \frac{1}{A} 0 = 0$ $(r \neq s)$ by (3) above. Hence $\| c_{rs} \| = \mathbf{I}$. Hence $\mathbf{AA}^{-1} = \mathbf{I}$ and similarly $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$, i.e. $\mathbf{A}^{-1}$ is one matrix such that $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. To show that it is the only such matrix, let $\mathbf{B}$ be any matrix such that $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$. Then:

$$\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{I} = \mathbf{A}^{-1}(\mathbf{AB}) = (\mathbf{A}^{-1}\mathbf{A})\mathbf{B} = \mathbf{IB} = \mathbf{B} \quad (\text{since } \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}).$$

Hence $\mathbf{B}$ must be $\mathbf{A}^{-1}$ and $\mathbf{A}^{-1}$ is unique. Q.E.D.

For non-singular square matrices $\mathbf{A}$, the basic property of inverses is: $\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. Other properties are easily derived:

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A}; \quad (\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}; \quad (\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'.$$

As a particular case $\mathbf{I}^{-1} = \mathbf{I}$, i.e. the unit matrix is such that it is the same matrix as its transpose and inverse $(\mathbf{I} = \mathbf{I}' = \mathbf{I}^{-1})$. The square matrix $\mathbf{A}$ with the property that $\mathbf{A}' = \mathbf{A}$ is a *symmetric* matrix, as noted at the end of 13.5. The square matrix $\mathbf{A}$ with the property that $\mathbf{A}^{-1} = \mathbf{A}'$ is also of considerable interest. Such a matrix is called *orthogonal* (13.9 Ex. 25). The unit matrix $\mathbf{I}$ happens to be both symmetric and orthogonal.

Given a square non-singular matrix of low order, we can always slog out the calculation of the inverse from the definition:

(i) Since $$\begin{Vmatrix} 1 & 2 \\ 0 & 1 \end{Vmatrix} \times \begin{Vmatrix} 1 & -2 \\ 0 & 1 \end{Vmatrix} = \begin{Vmatrix} 1 & 0 \\ 0 & 1 \end{Vmatrix}$$

by the multiplication rule, $\mathbf{A} = \begin{Vmatrix} 1 & 2 \\ 0 & 1 \end{Vmatrix}$ has inverse $\mathbf{A}^{-1} = \begin{Vmatrix} 1 & -2 \\ 0 & 1 \end{Vmatrix}$.

Checking: $A_{11}=1$, $A_{12}=0$, $A_{21}=-2$ and $A_{22}=1$ for the matrix **A** with determinant $A=1$. Hence $\mathbf{A}^{-1}=\left\|\begin{matrix}1 & -2\\ 0 & 1\end{matrix}\right\|$.

(ii) If $\mathbf{A}=\left\|\begin{matrix}1 & 2 & 3\\ 0 & 1 & 2\\ 0 & 0 & 1\end{matrix}\right\|$, then $\mathbf{A}^{-1}=\left\|\begin{matrix}1 & -2 & 1\\ 0 & 1 & -2\\ 0 & 0 & 1\end{matrix}\right\|$

$\mathbf{A}^{-1}$ is derived by calculating co-factors in the determinant $A=1$ as follows:

$$\begin{matrix}A_{11}=1 & A_{12}=0 & A_{13}=0\\ A_{21}=-2 & A_{22}=1 & A_{23}=0\\ A_{31}=1 & A_{32}=-2 & A_{33}=1.\end{matrix}$$

(iii) $\mathbf{A}=\left\|\begin{matrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{matrix}\right\|$ has inverse $\mathbf{A}^{-1}=\left\|\begin{matrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{matrix}\right\|=\mathbf{A}'$.

This is an example of an orthogonal matrix.

This method of calculation of the inverse $\mathbf{A}^{-1}$ of a given matrix **A**, from the definition, becomes very laborious as the order of **A** gets larger. Various techniques have been designed for inverting a matrix, for use on high-speed computers. The problem is important because of its application to linear transformations and systems of linear equations (13.8 below).

Consider the set of all square matrices of order $n \times n$ with elements drawn from a field F of scalars. Denote the set by $M_n(F)$. Since sums and scalar products of square matrices are appropriately defined, $M_n(F)$ is a vector space over F, the zero matrix being **O**, the square matrix of $n \times n$ zero elements. Further, any $n \times n$ matrix can be multiplied by any other $n \times n$ matrix, and $M_n(F)$ is closed under the non-commutative operation of multiplication. The associative and distributive properties hold and there is a unit matrix **I** of order $n \times n$ such that $\mathbf{AI}=\mathbf{IA}=\mathbf{A}$ for any $n \times n$ matrix **A**. Hence, $M_n(F)$ also has the structure of a ring, i.e. an additive group with a product operation satisfying the associative and distributive rules. On the other hand, $M_n(F)$ falls considerably short of a field. It lacks inverses, since only non-singular matrices have inverses and the singular matrices in $M_n(F)$ do not. Further, cancellation is not valid and there are matrix divisors of zero, as shown in an example in 13.5. So:

THEOREM: *The set of $M_n(F)$ all square matrices of a given order $n \times n$ is a vector space over the field of scalars and it has the structure of a ring.*

The product operation is associative and distributive, with a unit matrix **I**; *it lacks inverses and it has zero divisors.*

Square matrices, therefore, have most (though not all) of the desirable properties for three operations: sums, scalar products and products. $M_n(F)$ is called the *total matrix algebra.*

An improvement is achieved, as far as the product operation goes, when a more limited set of square matrices is taken: the set $L_n(F)$ of all non-singular square matrices of given order $n \times n$, a subset of $M_n(F)$. Since inverses now exist for all matrices of $L_n(F)$, and (consequently) cancellation becomes valid, the set is a non-commutative group under ×. It is called a *full linear group*:

Theorem: *The set $L_n(F)$ of all non-singular square matrices of given order $n \times n$ is a full linear group, i.e. a non-commutative group under the group operation of multiplication of matrices with a unit matrix* **I**.

Against this, there is a loss to record: the set $L_n(F)$ loses its standing as an additive group and hence as a vector space. In fact, it is not closed under addition since two non-singular matrices can add to a singular matrix, i.e. to a matrix outside $L_n(F)$. An example is given in 13.9 Ex. 14. It is, however, enough to quote the fact that a non-singular $\mathbf{A}$ and its non-singular negative $(-\mathbf{A})$ must add to the singular matrix $\mathbf{0}$.

13.7. The rank of a matrix. A matrix $\mathbf{A} = || a_{rs} ||$ of order $m \times n$ has elements from a field F of scalars. Consider the m rows of $\mathbf{A}$ as n-tuple vectors:

$$v_r = (a_{r1}, a_{r2}, \dots a_{rn}) \quad \text{for } r = 1, 2, \dots m.$$

A vector space V over F is generated by taking all linear combinations of the set $\{v_1, v_2, \dots v_m\}$ of m vectors, and let the dimension of this space be ρ. Since all vectors of V are linearly dependent on the set of m vectors, and since ρ is the largest number of linearly independent vectors in V, it follows that ρ cannot exceed m. A set of ρ linearly independent vectors is included within the set $\{v_1, v_2, \dots v_m\}$. Now $\rho = m$ is possible, implying that the row vectors $v_1, v_2, \dots v_m$ of $\mathbf{A}$ are linearly independent. Equally, $\rho < m$ is possible, implying that the m row vectors of $\mathbf{A}$ are themselves linearly dependent, i.e. linear combinations of some smaller set of ρ row vectors of $\mathbf{A}$. The integer ρ is called the rank of the matrix $\mathbf{A}$.

DEFINITION: *The* **rank** ρ *of the matrix* $\mathbf{A}$ *of order* $m \times n$ *is the dimension of the vector space of n-tuples generated by the rows of* $\mathbf{A}$. *Then* $\rho \leqslant m$, *where* $\rho = m$ *implies that the rows of* $\mathbf{A}$ *are linearly independent and where* $\rho < m$ *implies linear dependence among these rows.*

Hence, in general, there are ρ linearly independent rows in $\mathbf{A}$ and $m-\rho$ rows left over as dependent on them.*

Rank can be expressed in terms of determinants. The connection arises as a consequence of two basic results.

THEOREM: *If* $\mathbf{A} = \| a_{rs} \|$ *is of order* $n \times n$, *then the determinant* $A = 0$ *if and only if the rows of* $\mathbf{A}$ *are linearly dependent.*

Proof: if the rows $v_1, v_2, \ldots v_n$ as n-tuple vectors are linearly dependent, then one at least (say v_1) is a linear combination of the others:

$$v_1 = \lambda_2 v_2 + \lambda_3 v_3 + \ldots + \lambda_n v_n \quad \text{for some scalars } \lambda_2, \lambda_3, \ldots \lambda_n.$$

Writing the n-tuples as $v_r = (a_{r1}, a_{r2}, \ldots a_{rn})$ and separating off the sth components:

$$a_{1s} = \lambda_2 a_{2s} + \lambda_3 a_{3s} + \ldots + \lambda_n a_{ns} \quad \text{for } s = 1, 2, \ldots n \quad \ldots\ldots\ldots(1)$$

From (3) of 13.6 for rows:

$$A = \sum_{s=1}^{n} a_{1s} A_{1s} \quad \text{and} \quad \sum_{s=1}^{n} a_{2s} A_{1s} = \sum_{s=1}^{n} a_{3s} A_{1s} = \ldots = 0.$$

By (1):
$$A = \sum_{s=1}^{n} a_{1s} A_{1s} = \lambda_2 \sum_{s=1}^{n} a_{2s} A_{1s} + \lambda_3 \sum_{s=1}^{n} a_{3s} A_{1s} + \ldots = 0.$$

Hence, linear dependence of rows of $\mathbf{A}$ implies $A = 0$.

Take the converse proposition, and show that $A = 0$ implies linear dependence of rows of $\mathbf{A}$. If it happens that all the co-factors A_{rs} are zero, in producing $A = 0$, then it is clear that there is linear dependence in the rows of the co-factors and even more so in the rows of $\mathbf{A}$. To illustrate this, take the case $n = 3$:

$$A = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = 0; \; A_{11} = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} = 0; \; -A_{12} = \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} = 0; \ldots$$

* This development lacks symmetry and it is incomplete. There is no reason why it should be expressed (as here) in terms of rows as opposed to columns of $\mathbf{A}$. It can be shown (15.9 below) that the same result is obtained for columns as for rows. Let the n columns of $\mathbf{A}$ and their linear combinations form a vector space of dimension ρ', where $\rho' \leqslant n$. The basic result is that $\rho = \rho'$, i.e. the row rank and the column rank of a matrix are the same. Further, $\rho \leqslant$ smaller of m, n and $\mathbf{A}$ has ρ linearly independent and $m-\rho$ linearly dependent rows, and ρ linearly independent and $n-\rho$ linearly dependent columns. This result is reached, indirectly, at the end of the present section.

Now $A_{11}=0$ gives $a_{22}a_{33}-a_{23}a_{32}=0$ and either one or both rows of A_{11} consist of zero elements or one row of A_{11} is proportional to the other (i.e. $a_{22}=\lambda a_{32}$ and $a_{23}=\lambda a_{33}$ for some λ). If this holds for all the co-factors, then A must either have at least two rows (or columns) of zero elements or have each row proportional to another row. There is plenty of linear dependence among the rows of **A**. It remains to show that there is linear dependence even when the co-factors are not all zero. Suppose $A=0$ but one or more of the first column of co-factors (A_{11}, A_{21}, ... A_{n1}) are non-zero. From (3) of 13.6 for columns:

$$\sum_{r=1}^{n} a_{r1}A_{r1}=A=0 \quad \text{and} \quad \sum_{r=1}^{n} a_{r2}A_{r1}=\sum_{r=1}^{n} a_{r3}A_{r1}=\ldots=0.$$

Combining (a_{r1}, a_{r2}, ... a_{rn}) into the vector v_r:

$$\sum_{r=1}^{n} v_rA_{r1}=0 \quad \text{or} \quad A_{11}v_1+A_{21}v_2+\ldots+A_{n1}v_n=0$$

for multiples which are not all zero. Hence, if $A=0$, the rows v_1, v_2, ... v_n of **A** are linearly dependent. Q.E.D.

The second result deals with the various square sub-matrices which can be got from **A** of order $m\times n$ by suppressing certain rows and columns:

THEOREM: *If* $\mathbf{A}=\|a_{rs}\|$ *is of order* $m\times n$ *and rank* $\rho\leqslant m$, *then the largest square sub-matrix with a non-zero determinant is of order* $\rho\times\rho$.

Proof: suppose the largest such sub-matrix is **B** of order $k\times k$. Then $|\mathbf{B}|\neq 0$ and by the previous theorem the rows of **B** are linearly independent. Take the k rows of **A** which include **B**. These are also linearly independent; otherwise, if the k rows of **A** are linearly dependent, so are the rows of **B** (and this is not so). Since **A** has k linearly independent rows, its rank $\rho\geqslant k$. Next, take a square sub-matrix **C** of order $(k+1)\times(k+1)$ in **A**. Since **C** is bigger than **B**, we have $|\mathbf{C}|=0$ and by the previous theorem the rows of **C** are linearly dependent. The $k+1$ rows of **A** which include **C** are also linearly dependent; otherwise, if they are linearly independent, so are the rows of **C** (which is not so). Since **A** has $k+1$ linearly dependent rows, its rank $\rho<k+1$, i.e. $\rho\leqslant k$. So $\rho\geqslant k$ and $\rho\leqslant k$, i.e. $\rho=k$. Q.E.D.

The implication of this result is that there is at least one non-singular sub-matrix of order ρ in **A** but no non-singular sub-matrix of order greater than ρ. All square sub-matrices of **A** with more than

ρ rows and columns are singular. Hence, the rank ρ of **A** is *both* the largest number of linearly independent rows of **A** *and* the order of the largest non-singular sub-matrix in **A**. By the symmetry of this second property, it follows that ρ must arise equally from the rows and the columns of **A**, i.e. ρ is also the largest number of linearly independent columns of **A**. Further, it follows that $\rho \leqslant n$ as well as $\rho \leqslant m$.

For *square matrices*, the results can be summarised:

THEOREM: **A** *is of order* $n \times n$ *and of rank* $\rho \leqslant n$. *Then* $\rho = n$ *implies that* **A** *is non-singular and that all rows (or columns) are linearly independent;* $\rho < n$ *implies that* **A** *is singular and that* $n - \rho$ *of the rows (or columns) are dependent on the other* ρ.

For matrices which are not square, we can make the following statements. If **A** has $m < n$, i.e. fewer rows than columns, then ρ *must* be less than n and there are surplus columns in **A** ($n - \rho$ of them depending on the other ρ). If **A** has $n < m$, i.e. fewer columns than rows, then ρ *must* be less than m and there are surplus rows in **A** ($m - \rho$ of them depending on the other ρ).

13.8. Solution of linear equations. A general assault on the problem of 13.3 can now be made. Write:

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2 \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n = b_m \end{array}\right\} \quad \dots\dots\dots\dots\dots\dots(1)$$

as a system of m linear equations in n real variables $x_1, x_2, \dots x_n$. Once a solution of these equations is obtained, i.e. values of the x's in terms of the given b's (and the a's), then the parallel problem of inverting a linear transformation is also solved. Replace the b's in (1) by variables $y_1, y_1, \dots y_m$ and (1) becomes the linear transformation from the n-tuple $(x_1, x_2, \dots x_n)$ to the m-tuple $(y_1, y_2, \dots y_m)$. The solution of (1) becomes a set of relations giving the x's in terms of the y's (and the a's), i.e. the inverse linear transformation from the m-tuple $(y_1, y_2, \dots y_m)$ to the n-tuple $(x_1, x_2, \dots x_n)$.

The system (1) condenses neatly in the matrix notation. Write $\mathbf{A} = \| a_{rs} \|$ for the matrix of order $m \times n$ made up from the coefficients on the left of (1), and write $\mathbf{x}=\{x_1x_2 \ldots x_n\}$ for the column vector of the variable x's. The product of the $m \times n$ matrix **A** and the $n \times 1$ matrix **x** is a matrix of order $m \times 1$, i.e. another column vector comprising m components. By the inner product rule for matrix multiplication, this product **Ax** has m components which are precisely the expressions on the left of (1). Hence the column vector is the same as $\mathbf{b}=\{b_1b_2 \ldots b_m\}$, the column vector of given b's in (1). So:

$$\mathbf{Ax}=\mathbf{b} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

is the system of linear equations (1) in matrix notation.

Given the matrix **A** and the vector **b** in (2), the problem is to find the vector **x**. In general, **A** is of order $m \times n$ and of rank ρ.

Main case: $m=n=\rho$.

The system has as many equations as variables (both n). The matrix **A** is square (order $n \times n$) and, its rank being n, it is non-singular. The determinant $A \neq 0$ and the rows and columns of **A** are linearly independent. In this (very tidy) case, the system of equations has a unique solution vector **x**, i.e. the variables $x_1, x_2, \ldots x_n$ are given uniquely in terms of the given coefficients (the a's) and the given constants (the b's). The solution is obtained:

Since **A** is non-singular, write its inverse $\mathbf{A}^{-1}$. Pre-multiply each side of the matrix equation (2) by $\mathbf{A}^{-1}$: $\mathbf{A}^{-1}\mathbf{Ax}=\mathbf{A}^{-1}\mathbf{b}$. But

$$\mathbf{A}^{-1}\mathbf{Ax}=\mathbf{Ix}=\mathbf{x}.$$

Hence:

$$\mathbf{x}=\mathbf{A}^{-1}\mathbf{b} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Then (3) is the required solution of (2). To summarise:

THEOREM: *The solution of the linear equations* $\mathbf{Ax}=\mathbf{b}$, *where* **A** *is non-singular, is the unique vector* $\mathbf{x}=\mathbf{A}^{-1}\mathbf{b}$.

The solution in practice requires only some convenient methods of inverting the matrix **A**, suitable for calculation by hand, by desk machines or by computer. Such methods are available. When n is small, the equations can be solved by a process of eliminating variables in succession until only one is left, and there is also a fairly convenient formula for use when n is not large (13.9 Ex. 28). Two simple examples illustrate:

(i) $$\begin{aligned} x_1+2x_2+3x_3&=2 \\ x_2+2x_3&=1 \\ x_3&=1 \end{aligned} \quad \text{with } \mathbf{A}=\begin{Vmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{Vmatrix} \text{ and } \mathbf{A}^{-1}=\begin{Vmatrix} 1 & -2 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{Vmatrix}$$

as obtained in example (ii) of 13.6. The solution for $\mathbf{x}=\{x_1, x_2, x_3\}$ is:

$$\mathbf{x}=\mathbf{A}^{-1}\mathbf{b}=\begin{Vmatrix} 1 & -2 & 1 \\ 0 & 1 & -2 \\ 0 & 0 & 1 \end{Vmatrix}\times\begin{Vmatrix} 2 \\ 1 \\ 1 \end{Vmatrix}=\begin{Vmatrix} 1 \\ -1 \\ 1 \end{Vmatrix} \text{ i.e. } x_1=1, x_2=-1, x_3=1.$$

This can be checked by finding x_3, x_2 and x_1 in succession, taking the equations in order from the last to the first.

(ii)

$$\begin{aligned} -x_1+x_2+x_3&=a \\ x_1-x_2+x_3&=b \\ x_1+x_2-x_3&=c \end{aligned}$$

$$\text{with } \mathbf{A}=\begin{Vmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{Vmatrix} \text{ and } \mathbf{A}^{-1}=\begin{Vmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{Vmatrix}$$

as can be found from the definition of an inverse matrix. Hence:

$$\mathbf{x}=\mathbf{A}^{-1}\mathbf{b}=\begin{Vmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{Vmatrix}\times\begin{Vmatrix} a \\ b \\ c \end{Vmatrix}=\begin{Vmatrix} \frac{1}{2}(b+c) \\ \frac{1}{2}(c+a) \\ \frac{1}{2}(a+b) \end{Vmatrix}$$

i.e. $x_1=\frac{1}{2}(b+c)$, $x_2=\frac{1}{2}(c+a)$ and $x_3=\frac{1}{2}(a+b)$. This can be checked by eliminating x_3 and x_2 in succession and getting $x_1=\frac{1}{2}(b+c)$ in the end.

Degenerate cases: $\rho<m$ or $\rho<n$ or both.

In all cases, other than the main case, the rank ρ of **A** must be less than one or other or both of m and n. The cases are all lumped together with the label 'degenerate' since there is always some departure from the simplicity of solution which characterises the main case. The following assortment of cases are covered:

(*a*) $\rho<m=n$. There are as many equations as variables but the square matrix **A** of order $n\times n$ is singular, ρ being less than n. The number of linearly independent rows (or columns) of **A** is ρ and there are $n-\rho$ rows (or columns) left over as dependent on them.

(*b*) $\rho\leqslant m<n$. There are fewer equations than variables and the matrix **A** is not square. Here $\rho<n$ must hold, i.e. **A** has ρ linearly independent columns and $n-\rho$ left over as dependent on them. This is the case of 'surplus' variables.

(*c*) $\rho\leqslant n<m$. There are fewer variables than equations and **A** is not square. It must be that $\rho<m$ so that **A** has $m-\rho$ rows dependent on a set of ρ linearly independent rows. This is the case of 'surplus' equations.

The first question to decide is how the vector $\mathbf{b}=\{b_1 b_2 \ldots b_m\}$ of given constants relates to the matrix **A**. There are n columns in **A**, each an m-tuple vector, and $n-\rho$ of them are linearly dependent on the ρ others. Another m-tuple vector **b** is now added to the list. In the main case, where **A** has n columns of n-tuples, all linearly independent, the columns provide a basis for any set of n-tuples. Hence *any* additional n-tuple, such as **b**, is automatically a linear combination of the n columns of **A**. This is not so in the present degenerate cases, since the ρ linearly independent columns of m-tuples are not enough for a basis. The extra m-tuple **b** may or may not be linearly dependent on the columns of **A**.

If **b** is linearly dependent on the columns of **A**, written as the m-tuples $u_1, u_2, \ldots u_n$, then scalar multiples $\lambda_1, \lambda_2, \ldots \lambda_n$ exist so that $\lambda_1 u_1 + \lambda_2 u_2 + \ldots + \lambda_n u_n = \mathbf{b}$. Putting this relation between vectors in full detail:

$$\left.\begin{array}{l} a_{11}\lambda_1 + a_{12}\lambda_2 + \ldots + a_{1n}\lambda_n = b_1 \\ a_{21}\lambda_1 + a_{22}\lambda_2 + \ldots + a_{2n}\lambda_n = b_2 \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ a_{m1}\lambda_1 + a_{m2}\lambda_2 + \ldots + a_{mn}\lambda_n = b_m \end{array}\right\} \quad \ldots\ldots\ldots\ldots\ldots\ldots(4)$$

Comparing (4) with (1), we see that the system of linear equations does have at least the solution $x_1=\lambda_1$, $x_2=\lambda_2$, ... $x_n=\lambda_n$. On the other hand, if b is *not* linearly dependent on the columns of A, then *no* scalars $\lambda_1, \lambda_2, \ldots \lambda_n$ exist for (4) and the system of linear equations *must* be without solution. So:

THEOREM: *The system of linear equations* $\mathbf{Ax}=\mathbf{b}$ *has* **A** *of order* $m \times n$ *and rank* ρ, *where* $\rho < m$ *or* $\rho < n$ *or both; and it has* **b** *as a m-tuple vector not linearly dependent on the m-tuple columns of* **A**. *The system is then inconsistent and there is no solution.*

The inconsistency in the system is reflected in the fact that one or more of the equations are not consistent with one or more other equations. There is something wrong with the whole formulation. As an example:

$$\text{(iii)} \quad \begin{array}{r} x_1 - 2x_2 + 3x_3 = 1 \\ 2x_1 - x_2 + 4x_3 = 2 \\ x_1 + x_2 + x_3 = a \end{array} \quad \text{with} \quad \mathbf{A} = \left\Vert \begin{array}{ccc} 1 & -2 & 3 \\ 2 & -1 & 4 \\ 1 & 1 & 1 \end{array} \right\Vert \quad \text{of rank 2.}$$

The rank of **A** follows from the fact that $A = 0$ but $\begin{vmatrix} 1 & -2 \\ 2 & -1 \end{vmatrix} \neq 0.$

Subtract the first equation from the second: $x_1 + x_2 + x_3 = 1$. This is consistent with the third equation only if $a = 1$, in which case we continue to seek a solution. If a has any other value, e.g. $a = 0$, then the system of equations has no solution. In terms of linear dependence, the position is as follows. If u_1, u_2 and u_3 are the columns of **A**, then they are linearly dependent by virtue of the relation:

$$5u_1 - 2u_2 - 3u_3 = 0.$$

One of the columns is dependent on the other two. If the vector of constants is $\mathbf{b} = \{1 \ \ 2 \ \ 1\}$ corresponding to $a = 1$, then **b** is also linearly dependent on the columns of **A**. In fact, $\mathbf{b} = u_1 = 1 \times u_1 + 0 \times u_2 + 0 \times u_3$. But any other **b** such as $\{1 \ \ 2 \ \ 0\}$ is not linearly dependent in this way.

To continue, we suppose that **b** is checked to be linearly dependent on the columns of **A** and that we can look for a solution of the equations. The general position is that ρ is less than either or both of m and n. There are $m - \rho$ rows of **A** dependent on the set of ρ linearly independent rows. This means that there are $m - \rho$ surplus equations, dependent on or derivable from the others. These surplus equations are to be ignored. There are $n - \rho$ columns of **A** dependent on the set of ρ linearly independent columns. To correspond, there are ρ variables to use and $n - \rho$ surplus variables to which any values whatever can be assigned. Hence the system is to be viewed as giving only ρ variables (with the other $n - \rho$ assigned any values) by use of only ρ equations (the other $m - \rho$ being ignored). So:

THEOREM: *The system of linear equations* $\mathbf{Ax} = \mathbf{b}$ *has* **A** *of order* $m \times n$ *and rank* ρ, *where* $\rho < m$ *or* $\rho < n$ *or both; and it has* **b** *linearly dependent on the columns of* **A**. *The system is consistent and the solution is got by finding* ρ *of the variables from* ρ *of the equations. The other* $n - \rho$ *variables are assigned arbitrary values, and* $m - \rho$ *equations are ignored as derivable from the* ρ *equations.*

Of the three kinds of degenerate cases, (*a*) has an equal number (n) of equations and variables but $\rho < n$, so that there are both surplus variables and surplus equations. In (*b*), there are fewer equations than variables; there may be no surplus equations but there must be surplus variables. In (*c*), there are fewer variables than equations and, though there may not be surplus variables, there must be surplus equations. Each case is illustrated in the examples:

(iv) $x_1 - 2x_2 + 3x_3 = 1$ which is the consistent case of (iii) above.

$$\begin{array}{l} 2x_1 - x_2 + 4x_3 = 2 \\ x_1 + x_2 + x_3 = 1 \end{array}$$

The third equation follows from the other two, by subtracting the first from the second, and it can be ignored. Fix a value for x_3 and write the equations:

$$x_1 - 2x_2 = 1 - 3x_3 \quad \text{and} \quad 2x_1 - x_2 = 2(1 - 2x_3)$$

giving $x_1 = 1 - \frac{5}{3}x_3$ and $x_2 = \frac{2}{3}x_3$ in terms of x_3.

(v) $\begin{array}{l} 2x_1 + 4x_2 + x_3 = -1 \\ x_1 + 2x_2 + 2x_3 = 1 \end{array}$ with $\mathbf{A} = \begin{Vmatrix} 2 & 4 & 1 \\ 1 & 2 & 2 \end{Vmatrix}$ of rank 2.

There is a surplus variable, but no surplus equation. To see which variable to take as surplus, and to get an assigned value: take twice the second equation and subtract the first, giving $x_3 = 1$. Either equation then gives $x_1 + 2x_2 = -1$, i.e. either $x_1 = -(1 + 2x_2)$ with x_2 treated as surplus, or $x_2 = -\frac{1}{2}(1 + x_1)$ with x_1 so treated. So, if x_1 is assigned any values, the equations give: $x_2 = -\frac{1}{2}(1 + x_1)$ and $x_3 = 1$.

(vi) $\begin{array}{l} 2x_1 + x_2 = 0 \\ x_1 + 2x_2 = -3 \\ x_1 - x_2 = 3 \end{array}$ with $\mathbf{A} = \begin{Vmatrix} 2 & 1 \\ 1 & 2 \\ 1 & -1 \end{Vmatrix}$ of rank 2.

There is a surplus equation, any one of the three. For example, the second equation is derived by subtracting the last from the first. If it is ignored, then $2x_1 + x_2 = 0$ and $x_1 - x_2 = 3$ solve to give: $x_1 = 1$, $x_2 = -2$. This is the solution of the system of three equations, any one of them being ignored as implied by the others.

13.9. Exercises

1. Take the real numbers x as a vector space V over the field F of real numbers and show that it is a case of $V_1(F)$, the vectors being n-tuples ($n = 1$). Deduce that the space has dimension 1 and that the real number 1 can serve as a basis. Why is this different from example (i) of 13.2?

2. Show that, as a space $V_2(F)$ over the field F of real numbers, complex numbers have the pair 1 and i as a basis: $z = x \times 1 + y \times i$.

3. *Polynomials as vector spaces.* Extend example (iii) of 13.2, showing that the set of all cubics is a space $V_4(F)$, the set of all quartics a space $V_5(F)$, and so on. Then show that the set $F[x]$ of all polynomials is a vector space of infinite dimension.

4. Consider the 3×3 linear transformation of 13.3. Show that, by elimination of x_2 and x_3, x_1 can be found in terms of the y's, provided that $A \neq 0$ where:

$$A = a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}).$$

Then: $$x_1 = \frac{1}{A}\left\{y_1(a_{22}a_{33} - a_{23}a_{32}) - y_2(a_{21}a_{33} - a_{23}a_{31}) + y_3(a_{21}a_{32} - a_{22}a_{31})\right\}.$$

Find x_2 and x_3 similarly.

5. Apply the same method to the 2×3 linear transformation, (4) of 13.3, and show that

$$x_1 = \frac{1}{A}\left\{a_{22}y_1 - a_{12}y_2 + (a_{12}a_{23} - a_{13}a_{22})b_3\right\}$$

$$x_2 = \frac{1}{A}\left\{-a_{21}y_1 + a_{11}y_2 + (a_{13}a_{21} - a_{11}a_{23})b_3\right\}$$

provided that $x_3 = b_3$ is assigned and that $A = a_{11}a_{22} - a_{12}a_{21} \neq 0$. Examine similarly (5) of 13.3, showing that A must again be non-zero and that y_1, y_2 and y_3 must satisfy a certain relation.

*6. *Jacobians.* If $u_1, u_2, \dots u_m$ are each a function of a real variable x, the derivatives make up a row or column vector. Generalise to the Jacobian $J = \left\| \frac{\partial u_r}{\partial x_s} \right\|$, where each u is a function of n real variables. This matrix is named after Jacobi (1804–51).

7. Show that n inner products $\sum_{r=1}^{m} a_r b_{rs}$, for $s = 1, 2, \dots n$, can be written from a row vector $(a_1 a_2 \dots a_m)$ and a matrix $\| b_{rs} \|$ of order $m \times n$. Write the inner products for the same matrix and a column vector $\{c_1 c_2 \dots c_n\}$.

8. *Products of vectors and matrices.* Use Ex. 7 to show that $\mathbf{AB}$ exists where $\mathbf{A}$ is a row vector of order $1 \times m$ and $\mathbf{B}$ a matrix of order $m \times n$, and where $\mathbf{A}$ is a matrix of order $m \times n$ and $\mathbf{B}$ a column vector of order $n \times 1$. Hence interpret and express in full: $\mathbf{Ax}$ for $\mathbf{A} = \| a_{rs} \|$ $(r = 1, 2, \dots m;\ s = 1, 2, \dots n)$ and $\mathbf{x} = (x_1, x_2, \dots x_n)$.

*9. In n-dimensional Euclidean space (generalising 8.4), show that the length $|a|$ of a vector a is given by $|a|^2 = a \cdot a$, and that the angle α between vectors a and b is given by $\cos \alpha = \frac{a \cdot b}{|a|\,|b|}$.

10. *Negative and non-negative matrices.* Two notations can be used for 'greater than or equal to' applied to $\mathbf{A} = \| a_{rs} \|$: $\mathbf{A} \geqslant \mathbf{0}$ meaning $a_{rs} \geqslant 0$ all r and s (all $a_{rs} = 0$ not allowed); $\mathbf{A} \geqq \mathbf{0}$ meaning $a_{rs} \geqslant 0$ all r and s (all $a_{rs} = 0$ allowed). In both cases, $\mathbf{A}$ can be described as non-negative, i.e. whether $\mathbf{A} = \mathbf{0}$ is allowed or not. Show that it is still not true to say that non-negative $\mathbf{A} \geqq \mathbf{0}$ and negative $\mathbf{A} < \mathbf{0}$ cover all cases. (Note: some $a_{rs} > 0$ and some $a_{rs} < 0$ is a possibility.)

11. Show that $\left\| \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix} \right\| \left\| \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix} \right\| = \left\| \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix} \right\|$ and deduce that $\mathbf{I} = \mathbf{I}^2 = \mathbf{I}^3 = \dots$.

Show that $\mathbf{A}^r = \mathbf{A} \times \mathbf{A} \times \dots \times \mathbf{A}$ (r times) is defined if and only if $\mathbf{A}$ is square.

12. Suppose both $\mathbf{AB}$ and $\mathbf{BA}$ exist. Show that both products are square matrices but that they can be of different orders. Illustrate by multiplying

$\begin{Vmatrix} -2 & 1 & 0 \\ -3 & 0 & 1 \end{Vmatrix}$ and $\begin{Vmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{Vmatrix}$ to get a 2×2 or a 3×3 matrix according to the order of multiplication.

13. *Products of vectors.* Two vectors $x = (x_1, x_2, \dots x_n)$ and $y = (y_1, y_2, \dots y_n)$ have inner product $x \, . \, y = \sum_{s=1}^{n} x_s y_s$. In the matrix notation, show that x can be written $\mathbf{x}$ as a column vector and $\mathbf{x}'$ as a row vector (by transposing). Then show that the inner product $x \, . \, y = \mathbf{x}'\mathbf{y} = \mathbf{y}'\mathbf{x}$, as matrix products. On the other hand, show that $\mathbf{xy}'$ and $\mathbf{yx}'$ are $n \times n$ matrices, one the transpose of the other.

14. Illustrate that non-singular matrices can sum to singular matrices by showing that $\begin{Vmatrix} 2 & 1 \\ -2 & 0 \end{Vmatrix} + \begin{Vmatrix} -1 & -2 \\ 0 & 2 \end{Vmatrix} = \begin{Vmatrix} 1 & -1 \\ -2 & 2 \end{Vmatrix}$.

15. If $\mathbf{A} = \begin{Vmatrix} 2 & 1 \\ -1 & 0 \end{Vmatrix}$ and $\mathbf{B} = \begin{Vmatrix} 0 & -1 \\ 1 & 2 \end{Vmatrix}$, show that $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$ ($\mathbf{A}$ and $\mathbf{B}$ inverse) and that $\mathbf{A} + \mathbf{B} = 2\mathbf{AB}$.

16. If $\mathbf{A} = \| a_{rp} \|$, $\mathbf{B} = \| b_{pq} \|$, $\mathbf{C} = \| c_{qs} \|$ for $r = 1, 2, \dots m$, $p = 1, 2, \dots j$, $q = 1, 2, \dots k$ and $s = 1, 2, \dots n$, show that $\mathbf{AB}$ and $(\mathbf{AB})\mathbf{C}$ exist and specify their orders. By showing that $\sum_{p=1}^{j} a_{rp} b_{pq}$ is the (r, q)th element of $\mathbf{AB}$, deduce that the general element of $(\mathbf{AB})\mathbf{C}$ is $\sum_{q=1}^{k} \sum_{p=1}^{j} a_{rp} b_{pq} c_{rs}$. Hence show $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$.

17. If $\mathbf{A}$ is of order $m \times k$ and both $\mathbf{B}$ and $\mathbf{C}$ of order $k \times n$, show that $\mathbf{A}(\mathbf{B} + \mathbf{C})$ exists and equals $\mathbf{AB} + \mathbf{AC}$.

18. *Products with unit matrices.* Show that $\mathbf{AB} = \mathbf{IAB} = \mathbf{AIB} = \mathbf{ABI}$ provided only that $\mathbf{A}$ conforms with $\mathbf{B}$, but that $\mathbf{I}$ varies its order from one appearance to the next, except that the relations are quite unambiguous if $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{I}$ are all of order $n \times n$.

19. Under what conditions are the relations $\mathbf{AO} = \mathbf{OA} = \mathbf{O}$ valid for the product of a given matrix $\mathbf{A}$ and a zero matrix?

20. *Symmetric and skew-symmetric matrices.* $\| a_{rs} \|$ of order $n \times n$ is *symmetric* if $a_{rs} = a_{sr}$, and *skew-symmetric* if $a_{rs} = -a_{sr}$, all r and $s = 1, 2, \dots n$. Show that the leading diagonal can consist of any elements in the first case, but must comprise all zeros in the second. If $\mathbf{A}' = \mathbf{A}$, show that $\mathbf{A}$ is symmetric; if $\mathbf{A}' = -\mathbf{A}$ show that $\mathbf{A}$ is skew-symmetric.

*21. There are n partial derivatives $\dfrac{\partial u}{\partial x_r}$ for a function u of n real variables. Write second-order partial derivatives, illustrate from actual functions that they are symmetric and hence write the symmetric matrix $\mathbf{H} = \left\| \dfrac{\partial^2 u}{\partial x_r \, \partial x_s} \right\|$. This is called a *Hessian*.

22. If $\mathbf{A}$ is of order $n \times n$ and if $\mathbf{B} = \lambda \mathbf{A}$, show that $| \mathbf{B} | = \lambda^n \, | \mathbf{A} |$ and deduce that the singularity (or otherwise) of $\mathbf{A}$ is not affected by scalar multiplication.

23. *Division of matrices.* Show that non-singular matrices of the same order

can be divided but that division is non-commutative: $\mathbf{AB}^{-1}$ is one division of **A** by **B** and $\mathbf{B}^{-1}\mathbf{A}$ another, where in general $\mathbf{AB}^{-1} \neq \mathbf{B}^{-1}\mathbf{A}$.

24. If the scalars $\lambda_1, \lambda_2, \ldots \lambda_n$ are all non-zero, show that

$$\mathbf{A} = \left\| \begin{matrix} \lambda_1 & 0 \ldots 0 \\ 0 & \lambda_2 \ldots 0 \\ \cdot & \cdot \quad \cdot \quad \cdot \\ 0 & 0 \ldots \lambda_n \end{matrix} \right\| \text{ has } A = \lambda_1\lambda_2 \ldots \lambda_n \text{ and } \mathbf{A}^{-1} = \left\| \begin{matrix} 1/\lambda_1 & 0 \ldots 0 \\ 0 & 1/\lambda_2 \ldots 0 \\ \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \\ 0 & 0 \ldots 1/\lambda_n \end{matrix} \right\|.$$

*25. *Orthogonal matrices.* $\mathbf{A} = \| a_{rs} \|$ is *orthogonal* if $\mathbf{A}^{-1} = \mathbf{A}'$, i.e. if $\mathbf{AA}' = \mathbf{I}$. Write the (r, s)th element of $\mathbf{AA}'$ as an inner product and show that $\sum\limits_{t=1}^{n} a_{rt}^2 = 1$ for each r and that $\sum\limits_{t=1}^{n} a_{rt}a_{st} = 0$ for each r and s $(r \neq s)$. Interpret these relations and illustrate with $\mathbf{A} = \left\| \begin{matrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{matrix} \right\|$.

26. Establish that

$$\left\| \begin{matrix} 2 & 1 & 3 \\ 1 & 2 & 3 \\ 1 & -1 & 0 \end{matrix} \right\|, \quad \left\| \begin{matrix} 2 & 1 & 3 \\ 1 & 2 & 3 \end{matrix} \right\| \text{ and } \left\| \begin{matrix} 2 & 1 \\ 2 & 1 \\ 1 & -1 \end{matrix} \right\| \text{ are each of rank 2,}$$

but that both $\left\| \begin{matrix} 1 & -2 & 3 \\ -1 & 2 & -3 \end{matrix} \right\|$ and $\left\| \begin{matrix} 1 & 2 & 3 \\ 0 & 0 & 0 \end{matrix} \right\|$ are of rank 1.

*27. If **A** and **B** are of given order $m \times n$, show that the relation '**A** and **B** have the same rank' is an equivalence relation which serves to partition the set of all such matrices into equivalence classes according to rank.

28. *Cramer's rule.* If **A** is non-singular and of order $n \times n$, show that $\mathbf{Ax} = \mathbf{b}$ has solution $x_s = \sum\limits_{r=1}^{n} b_r A_{rs} \Big/ \sum\limits_{r=1}^{n} a_{rs} A_{rs}$ for $s = 1, 2, \ldots n$ where A_{rs} is the co-factor of a_{rs}. (To prove: multiply the n equations by $A_{1s}, A_{2s}, \ldots A_{ns}$ respectively, add and use (3) of 13.6.) Apply the rule to example (ii) of 13.8. The rule is named after Cramer (1704–52).

29. In the light of the results of 13.8, re-examine the solution of

$$a_{11}x_1 + a_{12}x_2 = b_1 \quad \text{and} \quad a_{21}x_1 + a_{22}x_2 = b_2.$$

Show that the degenerate cases are of two kinds:

(1) $\dfrac{a_{11}}{a_{21}} = \dfrac{a_{12}}{a_{22}} \neq \dfrac{b_1}{b_2}$ where the equations are inconsistent (no solution)

(2) $\dfrac{a_{11}}{a_{21}} = \dfrac{a_{12}}{a_{22}} = \dfrac{b_1}{b_2}$ where the equations are dependent and where a solution is obtained for x_1 (given x_2) or conversely. Interpret in terms of linear dependence between the columns of $\mathbf{A} = \left\| \begin{matrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{matrix} \right\|$ and the vector $\mathbf{b} = \left\| \begin{matrix} b_1 \\ b_2 \end{matrix} \right\|$.

30. *Homogeneous equations.* Consider $\mathbf{Ax} = \mathbf{0}$ as a set of n equations in n variables, where **A** of order $n \times n$ is given and where $\mathbf{x} = \{x_1x_2 \ldots x_n\}$. Assign a value of x_n, solve for $x_1, x_2, \ldots x_{n-1}$, using the results of 13.8 (degenerate cases) to show that the solution can be achieved if **A** is singular. Deduce that $\mathbf{Ax} = \mathbf{0}$ has solutions other than $x_1 = x_2 = \ldots = x_n = 0$ only if **A** is singular and that they are not unique.

31. Illustrate the result of Ex. 30 by showing that

$$x_1 - 2x_2 + 3x_3 = 0, \quad 2x_1 - x_2 + 4x_3 = 0 \quad \text{and} \quad x_1 + x_2 + x_3 = 0$$

has solutions $\frac{x_1}{-5} = \frac{x_2}{2} = \frac{x_3}{3}$. Deduce that there are unique *ratios* for the variables satisfying the three homogeneous equations in this case. Such a result is true in general when **A** is of order $n \times n$ and rank $n - 1$.

32. *Inversion of a linear transformation.* The linear transformation $\mathbf{y} = \mathbf{A}\mathbf{x}$, where **A** is of order $m \times n$ and rank ρ, is from the n-tuple **x** to the m-tuple **y**. Adapt the results of 13.8 to show that the inverse is $\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}$ if **A** is square and non-singular (main case) and that otherwise not all the variables can be used in inversion (degenerate cases). If $\rho < m$ and/or $\rho < n$, show that the transformation is consistent for inversion only if **y** is linearly dependent on the columns of **A**. Further, if $m < n$, show that some variables of **x** must be assigned before inversion; if $n < m$, show that the y's must satisfy one or more relations. See Ex. 5 above.

*33. *Orthogonal transformations.* The $n \times n$ linear transformation $\mathbf{y} = \mathbf{A}\mathbf{x}$ is orthogonal if **A** is orthogonal (Ex. 25). Show that $\mathbf{x} = \mathbf{A}'\mathbf{y}$ is the inverse, interpret and illustrate with $\mathbf{A} = \begin{Vmatrix} \cos\theta \sin\theta \\ -\sin\theta \cos\theta \end{Vmatrix}$. Orthogonal transformations of $V_n(F)$ into itself have the property that they preserve lengths; the 2 × 2 example here corresponds to a rotation of axes in two-dimensions.

*34. *Full linear groups.* Show that the linear transformations $\mathbf{y} = \mathbf{A}\mathbf{x}$ form a group under × (successive applications) for all non-singular matrices **A** of given order $n \times n$. Deduce that this group is isomorphic with $L_n(F)$, the group of non-singular matrices. Hence, algebraically, non-singular transformations and non-singular matrices are interchangeable concepts.

CHAPTER 14

LINEAR SYSTEMS

14.1 Linear algebraic systems. Linear algebra deals with sets which have the structure of a vector space and with two operations: sums and scalar products. A third operation, giving the products of elements of a set, may also be defined; but it is incidental and certainly not necessary. Concentrating on sums and scalar products, consider a *linear form*: $u = a_1x_1 + a_2x_2 + \dots + a_nx_n$. The a's are scalars, taken here as real numbers. The x's can be vectors (e.g. m-tuples) in which case u is a similar vector. However, take the x's in the present context as real variables so that u is also a real variable. The vector notation is always useful; for example, if a is the vector $(a_1, a_2, \dots a_n)$ and x the vector $(x_1, x_2, \dots x_n)$, then u is the inner product $a \,.\, x$.

What may be regarded as the essential feature of linearity in linear algebraic systems? We can develop an additive property which, at least, can lay claim to be of the essence of linearity. Though the property is perfectly general, the particular case of linear forms in two real variables is considered to simplify the exposition. Take first a pair of linear forms:

$$u = a_1x + b_1y \quad \text{and} \quad v = a_2x + b_2y \quad \dots\dots\dots\dots\dots\dots(1)$$

This is a linear transformation from pairs (x, y) to pairs (u, v). Then:

THEOREM: *If* $(x_1, y_1) \rightarrow (u_1, v_1)$ *and* $(x_2, y_2) \rightarrow (u_2, v_2)$ *under* (1), *then*

$$(\lambda_1x_1 + \lambda_2x_2, \lambda_1y_1 + \lambda_2y_2) \rightarrow (\lambda_1u_1 + \lambda_2u_2, \lambda_1v_1 + \lambda_2v_2)$$

for any scalars λ_1 *and* λ_2 *whatever.*

The additive property here is that, once two images are found under the transformation, other images follow as the sum of the given two, with any multiples λ_1 and λ_2 we care to take. The proof is immediate:

We know: $u_1 = a_1x_1 + b_1y_1$ and $u_2 = a_1x_2 + b_1y_2$.

So: $\lambda_1u_1 + \lambda_2u_2 = \lambda_1(a_1x_1 + b_1y_1) + \lambda_2(a_1x_2 + b_1y_2)$

$$=a_1(\lambda_1x_1+\lambda_2x_2)+b_1(\lambda_1y_1+\lambda_2y_2).$$

Similarly: $\lambda_1v_1+\lambda_2v_2=a_2(\lambda_1x_1+\lambda_2x_2)+b_2(\lambda_1y_1+\lambda_2y_2).$ Q.E.D.

Next consider a single equation, a linear form in two variables equated to zero. Write it in two ways, without and with an additive constant:

$$ax+by=0 \quad \text{.................................}(2)$$

$$ax+by+c=0 \quad \text{.............................}(3)$$

Here (2) is described as an equation in *homogeneous form* and (3) as one in *non-homogeneous form*. Moreover, given any equation (3), the corresponding homogeneous form (2) can always be written by dropping the constant c.

Suppose (x_1, y_1) and (x_2, y_2) are any two pairs satisfying the homogeneous form (2). Then it follows at once that the linear combination:

$$\lambda_1(x_1, y_1)+\lambda_2(x_2, y_2) \quad \text{i.e. the pair } (\lambda_1x_1+\lambda_2x_2,\ \lambda_1y_1+\lambda_2y_2)$$

also satisfies (2), for any λ_1 and λ_2 whatever. (The proof is as before.) In addition to these two pairs satisfying (2), suppose we have a particular pair $(\bar{x}, \bar{y})$ which satisfies (3). The striking result, then, is that

$$\lambda_1(x_1, y_1)+\lambda_2(x_2, y_2)+(\bar{x}, \bar{y})$$

i.e. the pair $(\lambda_1x_1+\lambda_2x_2+\bar{x},\ \lambda_1y_1+\lambda_2y_2+\bar{y})$

also satisfies the non-homogeneous form (3). The proof is again by substitution:

We know: $ax_1+by_1=0$, $ax_2+by_2=0$ and $a\bar{x}+b\bar{y}+c=0$.

Substitute $x=\lambda_1x_1+\lambda_2x_2+\bar{x}$ and $y=\lambda_1y_1+\lambda_2y_2+\bar{y}$ in (3):

$$a(\lambda_1x_1+\lambda_2x_2+\bar{x})+b(\lambda_1y_1+\lambda_2y_2+\bar{y})+c$$
$$=\lambda_1(ax_1+by_1)+\lambda_2(ax_2+by_2)+(a\bar{x}+b\bar{y}+c)=0.$$

The result obtained is:

THEOREM: *A solution of $ax+by+c=0$ is given by:*

$$x=\lambda_1x_1+\lambda_2x_2+\bar{x} \quad \textit{and} \quad y=\lambda_1y_1+\lambda_2y_2+\bar{y} \quad (\textit{for any } \lambda_1 \textit{ and } \lambda_2)$$

where (x_1, y_1) and (x_2, y_2) both satisfy $ax+by=0$ and where $(\bar{x}, \bar{y})$ satisfies $ax+by+c=0$.

The additive property here is that any two solutions of $ax+by=0$ (added with any multiples) and any solution of $ax+by+c=0$ can all be added together for a solution of the linear equation $ax+by+c=0$.*

The results obtained are, in fact, quite general and not confined to the particular case of two variables considered. Linear forms have an *additive property*: if particular solutions are found, then a general solution can be written by adding the particular solutions, with any multiples whatever.

There is one line of thought which is often followed to the conclusion that linearity is a very special and limited case. Suppose the real variable u depends on one or more other variables, e.g. u as a function of x, or as a function of x and y. Generally, we write $u=f(x)$ where the form of f is reflected in the shape of the curve which represents the relation graphically; or $u=f(x, y)$ and the form of f shows up in the shape of the corresponding three-dimensional surface. *Linear functions* such as $u=ax+b$ or $u=ax+by+c$ are indeed very special cases. They correspond to taking a line instead of a curve, a plane instead of a surface. The contrast is between the linear $u=ax+b$ (a line) and the quadratic $u=ax^2+bx+c$ (a parabola) or higher-order polynomials. These are all, in a sense, approximations to a general function $u=f(x)$ which is 'well-behaved' enough to have derivatives of all orders, i.e. approximations appropriate for a small neighbourhood of x around a particular value. These approximations are easily got by Taylor's series. For small x (around $x=0$), we have:

$$f(x)=f(0)+f'(0)x+f''(0)\frac{x^2}{2!}+\dots .$$

Approximate by neglecting x^2 and higher powers and

$$f(x)=f(0)+f'(0)x,$$

i.e. $u=ax+b$, with $a=f'(0)$ and $b=f(0)$. If x^3 and higher powers are neglected, then $u=ax^2+bx+c$, with $a=\frac{1}{2}f''(0)$, $b=f'(0)$ and $c=f(0)$. The linear function can, therefore, be regarded as the most severe approximation to a general function.

The linear relation is, however, not quite as limited as this. If u is a function of x, we can express $\log u$ as well as u in terms of $\log x$ as well as x. In a graphical representation, as an alternative to a graph on natural scales, we can use semi-logarithmic and logarithmic graphs (as employed in statistics). The linear function $u=ax+b$

* An interpretation: the relative position of two points on a line is required to determine the line's direction, and then one point is enough to fix its position.

(a and b given constants) is a line on natural scales. The function $u=be^{ax}$ gives: $\log u=ax+\log b$. This is linear on semi-logarithmic scales; $\log u$ is a linear function of x and $u=be^{ax}$ is shown by a line when $\log u$ is plotted against x. Further, the function $u=bx^a$ gives: $\log u=a\log x+\log b$. This is linear on logarithmic scales; $\log u$ is linear in $\log x$ and $u=bx^a$ is a line when $\log u$ is plotted against $\log x$. Consequently, all the functions $u=ax+b$, $u=be^{ax}$ and $u=bx^a$ can be regarded as linear, and between them they cover a considerable range. In particular, while $u=ax+b$ expresses a constant absolute rate (a) of growth of u with respect to x, $u=be^{ax}$ is a constant proportionate rate (a) of growth (12.3 above). They are both linear, as shown in the constant rate of growth.

It remains to explore an extension of the idea of linearity, one which is suggested by the last remark. We turn from a linear form in one or more variables to a consideration of linearity in the growth of one variable in relation to another, and in particular to variation over time. The emphasis is on the time-path of a dynamic variable as opposed to a static value.

14.2. Linear differential equations. The relation $y=be^{ax}$ represents growth at a constant proportionate rate, e.g. it shows the growth of a sum of money at various times (x years) when interest is compounded continuously at $100a$ per cent per year. Consider the derivation of the relation. We are given only one fact: y grows at the given proportionate rate a. If D is the operator for a derivative, the given fact is:

$$\frac{1}{y}Dy=a \quad \text{or} \quad D\log y=a \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

The relation between y and x is then to be found as an anti-derivative or integral:

$$\log y=\int a\,dx+\text{constant}=ax+\text{constant}$$

giving $$y=e^{ax+\text{constant}}=e^{\text{constant}}\,e^{ax}$$

i.e. $$y=be^{ax} \quad (b\text{ constant}) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

The 'b' here is an arbitrary constant, but it has an interpretation as the value of y when $x=0$. If we are given an extra fact, that $y=y_0$ when $x=0$, then the relation is unambiguous:

$$y=y_0e^{ax} \quad (y_0\text{ initial value of }y)\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

For example, this shows the amount £y of an initial sum £y_0 at the end of x years at continuously compounded interest of $100a$ per cent per year.

In reviewing this problem, we note that (1) is a 'differential equation', an equation including the derivative Dy as well as y. We find a 'solution' by integration. In this case, it is (2), where b is some 'arbitrary' constant. Or, it is (3), where an additional fact is known, the 'initial' value y_0 at $x=0$. The problem can be generalised.

We seek to express a variable y as a function of another variable x. We are given only a relation between x, y and various derivatives:

DEFINITION: *An (ordinary)* **differential equation** *is some relation:*

$$F(x, y, Dy, D^2y, \dots D^ny)=0$$

and its **order** *n is that of the highest derivative D^ny included.*

Further, attach the label 'linear' to a particular case:

DEFINITION: *A differential equation of order n is* **linear** *if it is of the form:*

$$D^ny+a_1(x)D^{n-1}y+\dots+a_{n-1}(x)Dy+a_n(x)y=\phi(x)$$

where $a_1(x)$, ... $a_{n-1}(x)$, $a_n(x)$ are specified functions. It is **linear with constant coefficients** *if it is:*

$$D^ny+a_1D^{n-1}y+\dots+a_{n-1}Dy+a_ny=\phi(x)$$

where a_1, ... a_{n-1}, a_n are specified constants.

The notation here is in terms of the operator D. Differential equations are often written with the alternative notation $\dfrac{d}{dx}$ for D; the linear differential equation with constant coefficients is then:

$$\frac{d^ny}{dx^n}+a_1\frac{d^{n-1}y}{dx^{n-1}}+\dots+a_{n-1}\frac{dy}{dx}+a_ny=\phi(x).$$

To solve a differential equation is to find the form of the function $y=f(x)$ which satisfies it. One simple case is illuminating, the first order linear differential equation in which the term in y is absent: $Dy=\phi(x)$. The solution is known to be: $y=\int\phi(x)\,dx+\text{constant}$. In a sense, solving a differential equation is a generalised form of finding an integral or anti-derivative. For this reason, the solution is often termed the integral of the differential equation.

In exploring the general nature of the solution, we can first enquire why the *first* order equation $Dy=\phi(x)$ has just *one* arbitrary constant.

If we are given $y=f(x;\,A)$, including an arbitrary constant A, then $Dy=f'(x;\,A)$. The constant A can be eliminated between $y=f(x;\,A)$ and $Dy=f'(x;\,A)$ to give some relation between x, y and Dy, i.e. to give a first order differential equation. One derivative gets rid of an arbitrary constant; integration brings it back again. Similarly, given $y=f(x;\,A,\,B)$ with two arbitrary constants, then A and B can be eliminated between y, $Dy=f'(x;\,A,\,B)$ and $D^2y=f''(x;\,A,\,B)$, and the result is a second order differential equation. In general, if n arbitrary constants are included in the relation of y to x, then they are eliminated in writing some nth order differential equation. Conversely, we expect to find n arbitrary constants in the general solution of a differential equation of order n.

This has an important consequence; there must be all kinds of particular functions satisfying a differential equation. For example, suppose that the *general* solution of a second-order equation is $y=Af_1(x)+Bf_2(x)$, where A and B are arbitrary constants. Then $y=f_1(x)$ is a *particular* solution ($A=1$, $B=0$) and $y=f(x_2)$ is another ($A=0$, $B=1$); but so are $y=f_1(x)+f_2(x)$ and many others. It is of no use checking (say) that $y=f_1(x)$ satisfies the equation and leaving it at that. In this way *a* solution is obtained, but not the general solution with the appropriate arbitrary constants.

The next step is to find the form of the general solution of a linear differential equation. Suppose the equation is of order n:

$$D^n y + a_1 D^{n-1} y + \dots + a_{n-1} Dy + a_n y = \phi(x) \quad \dots\dots\dots\dots\dots(4)$$

and write the corresponding *homogeneous form*:

$$D^n y + a_1 D^{n-1} y + \dots + a_{n-1} Dy + a_n y = 0 \quad \dots\dots\dots\dots\dots(5)$$

where the a's are given functions of x (or constants in the particular case). It can now be shown that the same additive property holds for linear differential equations as for linear algebraic equations (14.1). This is the justification for the term linear. Equations like (4) or (5) form a linear system.

First, suppose that $y=y_1(x)$ and $y=y_2(x)$ are known (e.g. by checking them in the equation) to satisfy the homogeneous form (5). Then $y=A_1y_1(x)+A_2y_2(x)$ also satisfies (5), and for any constants A_1 and A_2. For, on substituting:

$$D^n(A_1y_1+A_2y_2)+a_1D^{n-1}(A_1y_1+A_2y_2)+\dots$$
$$+a_{n-1}D(A_1y_1+A_2y_2)+a_n(A_1y_1+A_2y_2)$$

$$=A_1(D^ny_1+a_1D^{n-1}y_1+\dots$$
$$+a_{n-1}Dy_1+a_ny_1)\ +A_2(D^ny_2+a_1D^{n-1}y_2+\dots+a_{n-1}Dy_2+y_2)$$
$$=0 \quad \text{since } y_1 \text{ and } y_2 \text{ are solutions.}$$

Next, suppose that $y=\bar{y}(x)$ is known to satisfy the non-homogeneous form (4) and $y_1(x)$ is some solution of the homogeneous form (5). Then $y=y_1(x)+\bar{y}(x)$ also satisfies (4). For on substituting:

$$D^n(y_1+\bar{y})+a_1D^{n-1}(y_1+\bar{y})+\dots+a_{n-1}D(y_1+\bar{y})+a_n(y+\bar{y})$$
$$=(D^ny_1+a_1D^{n-1}y_1+\dots+a_{n-1}Dy_1+a_ny_1)+(D^n\bar{y}+a_1D^{n-1}\bar{y}+\dots$$
$$+a_{n-1}D\bar{y}+a_n\bar{y})$$
$$=\phi(x) \quad \text{since the first bracket is zero and the second } \phi(x).$$

The two results can be combined and developed to establish:

THEOREM: *The general solution of a linear differential equation of order n is:*

$$y=A_1y_1(x)+A_2y_2(x)+\dots+A_ny_n(x)+\bar{y}(x) \quad \dots\dots\dots\dots(6)$$

where $y_1(x)$, $y_2(x)$, ... $y_n(x)$ are n different particular solutions of the homogeneous form, where $\bar{y}(x)$ is any particular solution of the non-homogeneous form, and where A_1, A_2, ... A_n are arbitrary constants.

Here $\bar{y}(x)$ is called the *particular integral* and the linear combination of the functions $y_1(x)$, $y_2(x)$, ... $y_n(x)$ is the *complementary function.* It is to be stressed that the n functions in the complementary function must be all *different* and in the genuine sense, excluding (e.g.) functions which differ only by constants. To ensure a general solution, n particular and different solutions of the homogeneous form must be obtained and the linear combination of them written; the addition of any solution of the original differential equation completes the solution.

Given only the linear differential equation itself, there is not a unique solution. There are n arbitrary constants to be assigned; by allotting various values to them, we get a range of particular solutions. Suppose, however, that something else is given, i.e. n initial values such as:

$$y=y_0,\ Dy=y_0',\ D^2y=y_0'',\ \dots\ D^{n-1}y=y_0^{(n-1)} \quad \text{at } x=0.$$

Then, on substituting the general solution (6), n equations are obtained in the n arbitrary constants and the n initial values

$$y_0,\ y_0',\ y_0'',\ \dots\ y_0^{(n-1)}.$$

From these, in general, the arbitrary constants can be expressed in terms of the initial values. The solution (6) then becomes unique, given the initial values. This is the usual form in which the problem is presented. The solution of a linear differential equation of order n, subject to n initial conditions, is unique. It is (6) with the values of $A_1, A_2, \ldots A_n$ expressed in terms of the given initial conditions. For example, the differential equation (1), or $Dy-ay=0$, is linear and homogeneous, of first order, with constant coefficients. Its solution is (2) where b is an arbitrary constant. Given the initial condition $y=y_0$ when $x=0$, the solution is (3) and unique.

14.3. Solution of linear differential equations. The practical techniques for solving specific differential equations are many and various and they are not to be pursued here. Even for linear equations, it is one thing to write the solution in the form (6) of 14.2 but it is quite another matter to spell it out. The problem has been simplified, to the extent of transforming it into the problem of finding, first, the n particular constituents of the complementary function, and then the particular integral. This residual problem is far from easy. This is particularly so when the linear equation does not have constant coefficients. Indeed, in this case, it sometimes happens that no particular solution of the homogeneous form can be found in terms of known functions, i.e. that the equation defines a new function. This is a natural extension of the method of defining new functions by means of an integral (a first-order linear differential equation) as pursued in Chapter 12. An illustration is given in 14.9 Ex. 5.

The case analysed here is the particular one of a linear differential equation with constant coefficients. It is possible in this type of equation to complete the solution in general terms, but only for the complementary function with n arbitrary constants. The question of finding the particular integral is left open; at best it is something of a hit-or-miss affair. Enough is done here to establish two things of general interest and practical importance. One is that the solution of a linear differential equation (with constant coefficients) is not only of the same additive kind as that of a linear algebraic equation, but is also obtained in practice by solving an algebraic equation. The first step in finding the complementary function of a linear differential equation is to reduce the equation to an algebraic (polynomial) equation. Since we can always solve the latter, in one way or another in practice, we can also solve the differential equation.

The other point is that the solutions of linear differential equations quite often, indeed usually, include oscillatory components of the form of the circular functions of 12.5. This is so of equations of as low an order as the second. Oscillatory movements appear in many problems in the natural and social sciences. The problems are framed by specifying the differential equations satisfied by the variables and the solution of the equations shows the oscillatory nature of the movements of the variables.

The *first-order* differential equation with constant coefficients is easily handled. It can be written: $Dy + ay = \phi(x)$. To obtain the complementary function, which has only one term, write the homogeneous form: $Dy = -ay$. So:

$$D \log y = \frac{1}{y} Dy = -a$$

i.e. $$\log y = -ax + \text{constant}$$

and $$y = Ae^{-ax}$$

is the complementary function with its single arbitrary constant A. To complete, we must find a particular integral $\bar{y}(x)$ of the original equation $Dy + ay = \phi(x)$. This is usually a matter of trial and error, according to the form of $\phi(x)$. The general solution of the original equation is then:

$$y = Ae^{-ax} + \bar{y}(x).$$

In the particular case $a = 0$, the complementary function is simply $y = A$. To this, the particular integral $\bar{y}(x)$ is to be added. Directly, the equation is: $Dy = \phi(x)$ and the solution is $y = \int \phi(x)\,dx + A$. The particular integral is just the indefinite integral of $\phi(x)$. Another example illustrates:

(i) $Dy + y = e^x$, with complementary function $y = Ae^{-x}$.

As a guess, try $\bar{y} = ke^x$ as a particular integral. Substituting $\bar{y} = D\bar{y} = ke^x$:

$$D\bar{y} + \bar{y} = 2ke^x = e^x \quad \text{if } k = \tfrac{1}{2}.$$

Hence $\bar{y} = \frac{1}{2}e^x$ and the complete solution is $y = Ae^{-x} + \frac{1}{2}e^x$.

The *second-order* differential equation with constant coefficients and its corresponding homogeneous form are:

$$D^2y + aDy + by = \phi(x) \qquad \text{.........................(1)}$$

and $$D^2y + aDy + by = 0 \quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

The particular integral $y = \bar{y}(x)$ is to be obtained, somehow, from (1). The complementary function is $y = A_1y_1(x) + A_2y_2(x)$, where y_1 and y_2 are two different particular solutions of (2). It does not matter how we obtain y_1 and y_2, as long as we get them. A trick suggested by the solution of the first-order equation, and adopted with no apology, is to try $y = e^{\lambda x}$ as a solution of (2) and to see whether we can find two different λ's. In (2), substitute $y = e^{\lambda x}$, $Dy = \lambda e^{\lambda x}$ and $D^2y = \lambda^2 e^{\lambda x}$:

$$(\lambda^2 + a\lambda + b)e^{\lambda x} = 0.$$

Cancel $e^{\lambda x} > 0$ (for all x) and get the *auxiliary equation*:

$$\lambda^2 + a\lambda + b = 0 \quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

with two values of λ:

$$\lambda_1, \lambda_2 = \tfrac{1}{2}\{-a \pm \sqrt{(a^2 - 4b)}\}. \quad\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

The complementary function, the solution of (2), is then:

$$y = A_1e^{\lambda_1 x} + A_1e^{\lambda_2 x} \quad\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

provided only that the values of λ given by (4) are different. The complete solution of (1) is then written by the addition of the particular integral $\bar{y}(x)$ to (5).

This remarkable result implies that the solution (complementary function) of the differential equation (2) is got simply by replacing it by the polynomial (algebraic) equation (3). The second-order (2) gives the quadratic (3), the coefficients being the same. The result generalises to a differential equation of any linear order with constant coefficients; if (2) is of order n, then (3) is a polynomial equation of nth degree. The complementary function (5) also contains n terms, corresponding to the n roots of the auxiliary equation. There is nothing more to be said in general. In practice, having reduced the differential equation to the algebraic auxiliary equation, we concentrate on getting the n roots we know the auxiliary equation has. The problems left to be tackled are problems of detail.

Pursuing the detail of the second-order equation, we distinguish the three cases of the roots (4) of the quadratic auxiliary equation.

Case: $a^2 > 4b$. The roots λ_1 and λ_2 are real and distinct. The solution of (1) is:

$$y = A_1e^{\lambda_1 x} + A_2e^{\lambda_2 x} + \bar{y}(x).$$

This is very similar to the solution of the first-order equation. The only difference is that there are two exponential terms instead of one. These terms are increasing or decreasing according as the λ's are positive or negative. For example, if $\lambda_1 > 0$ or $\lambda_2 > 0$, then $y \to \infty$; if $\lambda_1 < 0$ and $\lambda_2 < 0$, then y approaches $\bar{y}$ as $x \to \infty$.

Case: $a^2 = 4b$. The roots λ_1 and λ_2 are real and equal. Here there is a residual difficulty; since there is only one λ, there is (as yet) only one term in (5) and we need two. The single $\lambda = -\frac{1}{2}a = -\sqrt{b}$, given by the coefficients of the differential equation. The latter can, therefore, be written in homogeneous form:

$$D^2y - 2\lambda Dy + \lambda^2 y = 0.$$

Another trick is needed to complete the solution. The one which works is to try $y = xe^{\lambda x}$. On substituting $y = xe^{\lambda x}$, $Dy = (1 + \lambda x)e^{\lambda x}$ and $D^2y = \lambda(2 + \lambda x)e^{\lambda x}$:

$$D^2y - 2\lambda Dy + \lambda^2 y = \{\lambda(2 + \lambda x) - 2\lambda(1 + \lambda x) + \lambda^2 x\}e^{\lambda x} = 0.$$

Hence, the second particular solution is $y = xe^{\lambda x}$ to add to the first, $y = e^{\lambda x}$. The solution of (1) is:

$$y = (A_1 + A_2 x)e^{\lambda x} + \bar{y}(x)$$

and this does not differ substantially from that of the first case.

Case: $a^2 < 4b$. The roots λ_1 and λ_2 are conjugate complex. This is the case where the solution of (1) is oscillatory. It merits separate examination (14.4).

Meanwhile, the following examples illustrate:

(ii) $D^2y + 3Dy + 2y = 0$ with auxiliary equation

$$\lambda^2 + 3\lambda + 2 = (\lambda + 1)(\lambda + 2) = 0.$$

The solution is: $y = A_1 e^{-x} + A_2 e^{-2x} = (A_1 + A_2 e^{-x})e^{-x} \to 0$ as $x \to \infty$. The equation $D^2y + 2Dy + y = 0$, with auxiliary equation

$$\lambda^2 + 2\lambda + 1 = (\lambda + 1)^2 = 0,$$

has a rather similar solution: $y = (A_1 + A_2 x)e^{-x} \to 0$ as $x \to \infty$.

(iii) $D^2y - y + x = 0$. The complementary function is the solution of $D^2y - y = 0$, with auxiliary equation $\lambda^2 - 1 = 0$, i.e. it is $y = A_1 e^x + A_2 e^{-x}$. As a particular integral, try $y = kx$. On substitution: $-kx + x = 0$, i.e. $k = 1$. Hence the solution of the equation is: $y = A_1 e^x + A_2 e^{-x} + x$.

If the problem is framed: find the solution of $D^2y-y+x=0$, subject to initial conditions $y=y_0$, $Dy=0$ at $x=0$, then

$$y_0=A_1+A_2 \quad \text{and} \quad 0=A_1-A_2 \quad \text{(from } y \text{ and } Dy \text{ at } x=0).$$

Hence, $A_1=\frac{1}{2}y_0$ and $A_2=\frac{1}{2}y_0$. The unique solution is:

$$y=\tfrac{1}{2}y_0(e^x+e^{-x})+x.$$

A more general problem arises when there are several variables $(y, z, u, \ldots)$, each a function of x, and subject to several simultaneous differential equations. As a simple case, which can be generalised a little (14.9 Ex. 10), consider a pair of first-order, linear, homogeneous differential equations, in two variables, y and z, each a function of x:

$$Dy=a_{11}y+a_{12}z \quad \text{and} \quad Dz=a_{21}y+a_{22}z \quad \ldots\ldots\ldots\ldots\ldots(6)$$

where the a's are given constants making up a matrix $\mathbf{A}=\|\, a_{rs}\,\|$, and giving a determinant $A=|\,a_{rs}\,|=a_{11}a_{22}-a_{12}a_{21}$, for r and $s=1, 2$. The variable z can be eliminated from (6) by using the first equation:

$$z=\frac{1}{a_{12}}\Big(Dy-a_{11}y\Big) \quad \text{and so} \quad Dz=\frac{1}{a_{12}}\Big(D^2y-a_{11}Dy\Big)$$

and by substituting in the second equation:

$$\frac{1}{a_{12}}\Big(D^2y-a_{11}Dy\Big)=a_{21}y+\frac{a_{22}}{a_{12}}\Big(Dy-a_{11}y\Big) \quad (a_{12}\neq 0).$$

Hence y satisfies the *second-order* linear differential equation:

$$D^2y-(a_{11}+a_{22})Dy+Ay=0 \quad (a_{12}\neq 0) \quad \ldots\ldots\ldots\ldots\ldots(7)$$

Equally, by using the second equation to give y and Dy and by substituting in the first equation, we find that z satisfies precisely the same equation:

$$D^2z-(a_{11}+a_{22})Dz+Az=0 \quad (a_{21}\neq 0).$$

This is obvious enough from the symmetric way in which the a's appear in (7). Hence the movements of y and z as x varies are identical, apart from a multiplicative constant (k) which leaves (7) unchanged. The solution of (7) is:

$$y=A_1e^{\lambda_1x}+A_2e^{\lambda_2x} \quad (A_1 \text{ and } A_2 \text{ arbitrary})$$

where λ_1 and λ_2 are the roots of $\lambda^2-(a_{11}+a_{22})\lambda+A=0$. The first equation then gives: $z=\dfrac{1}{a_{12}}\,(Dy-a_{11}y)$. On substitution for y:

$$z = k_1 A_1 e^{\lambda_1 x} + k_2 A_2 e^{\lambda_2 x} \quad \left(k_1 = \frac{\lambda_1 - a_{11}}{a_{12}} \quad \text{and} \quad k_2 = \frac{\lambda_2 - a_{11}}{a_{12}}\right).$$

The fact that y and z follow essentially the same path (e.g. over time x) is a consequence of the assumption at the outset of the linear forms (6).

The paths of y and z are similar exponential growths (or declines) if the auxiliary equation of (7) has real roots, i.e. if $A \leqslant \frac{1}{4}(a_{11} + a_{22})^2$. Otherwise, if the auxiliary equation has conjugate complex roots, then y and z have similar oscillatory paths.

As a particular case of (6), suppose that the matrix **A** is singular, i.e. that $A = a_{11}a_{22} - a_{12}a_{21} = 0$. This means that the ratio $a_{11} : a_{21}$ is the same as the ratio $a_{12} : a_{22}$. The auxiliary equation of (7) gives $\lambda_1 = 0$ and $\lambda_2 = (a_{11} + a_{22}) = \mu$ (say). The solution is of the form:

$$\left.\begin{aligned} y &= A + Be^{\mu x} \\ \text{and} \qquad z &= k_1 A + k_2 B e^{\mu x} \end{aligned}\right\} \quad A \text{ and } B \text{ arbitrary.}$$

As $x \to \infty$ (e.g. as time goes on), y and z both behave like $e^{\mu x}$ and $\frac{z}{y} \to k_2$ (constant).

14.4. Oscillatory movements. We return to the linear and homogeneous differential equation $D^2 y + aDy + b = 0$ in the case $a^2 < 4b$ where the auxiliary equation $\lambda^2 + a\lambda + b = 0$ has conjugate complex roots $\frac{1}{2}\{-a \pm i\sqrt{(4b - a^2)}\}$. A convenient notational change for the structural constants (a and b) of the equation is in order. Write the conjugate complex roots as $\alpha \pm i\omega$, where α and ω are constants given in terms of a and b, i.e. α and ω are (alternative) structural constants of the equation. The relations between the alternative constants are: $\alpha = -\frac{1}{2}a$ and $\omega = \frac{1}{2}\sqrt{(4b - a^2)}$ Hence:

$$a = -2\alpha \quad \text{and} \quad b = \alpha^2 + \omega^2.$$

The differential equation can be written in terms of the new structural constants:

$$D^2 y - 2\alpha D y + (\alpha^2 + \omega^2) y = 0$$

and the solution can be written

$$y = A_1 e^{(\alpha + i\omega)x} + A_2 e^{(\alpha - i\omega)x} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

for arbitrary constants A_1 and A_2 which can also be complex values.

This solution, though neat, can be developed into alternative forms, which are of greater use in practice, and which do not involve complex values. The differential equation is in a *real* variable y and it has *real* coefficients involving α and ω. The complex values in (1) are merely an intermediate step to a real solution.

The results of 12.6 and 12.7 come into play in this development. Write (1) as:

$$y = e^{\alpha x}(A_1 e^{i\omega x} + A_2 e^{-i\omega x})$$

$$= e^{\alpha x}\{A_1(\cos \omega x + i \sin \omega x) + A_2(\cos \omega x - i \sin \omega x)\}$$

i.e. $$y = e^{\alpha x}(B_1 \cos \omega x + B_2 \sin \omega x) \quad \ldots\ldots\ldots\ldots(2)$$

where $B_1 = A_1 + A_2$ and $B_2 = i(A_1 - A_2)$ are also arbitrary constants. We now have y in its proper real form so that B_1 and B_2 must be real constants. (It follows that the original constants A_1 and A_2 are conjugate complex values.) The solution (2) shows the real path of y, as x changes, more explicitly than the equivalent (1).

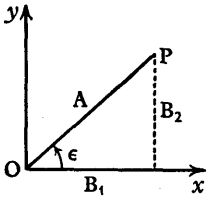


FIG. 14 .

A further shift in the notation for the arbitrary constants can be made. The nature of the change is seen in Fig. 14.4a; it is, in effect, a switch from the Cartesian co-ordinates (B_1, B_2) of a point P to the polar co-ordinates (A, ϵ). Write:

$$B_1 = A \cos \epsilon \quad \text{and} \quad B_2 = A \sin \epsilon$$

which is the same thing as writing:

$$A = \surd(B_1^2 + B_2^2) \quad \text{and} \quad \epsilon = \tan^{-1}(B_2/B_1)$$

(see Appendix A.9).

Hence, since B_1 and B_2 are arbitrary real constants, so are A and ϵ. Then (2) becomes:

$$y = e^{\alpha x}(A \cos \omega x \cos \epsilon + A \sin \omega x \sin \epsilon).$$

By use of the addition formula, (2) of 12.5:

$$y = Ae^{\alpha x} \cos(\omega x - \epsilon) \quad \ldots\ldots\ldots\ldots(3)$$

which is the most concise and convenient form for the solution of the differential equation. There are four parameters in (3) and it is important to distinguish between them. Two of them, α and ω, are given by the structure of the differential equation taken; they correspond to the inherent or structural variation of y. The other two, A and ϵ, are arbitrary constants to be given by initial conditions;

they correspond to the 'accidental' variation of y, arising because it starts off in a particular way.

The function (3) is a generalised circular function, called a *sinusoidal function*, taking its shape (when represented graphically) from the regular and symmetric oscillation of the cosine function (12.7 above). The dependence of the oscillation on the four parameters needs careful examination. Write (3) in two parts:

$$y=uv \quad \text{where} \quad u=Ae^{\alpha x} \quad \text{and} \quad v=\cos(\omega x-\epsilon).$$

Then v is the oscillatory term and u simply serves to amplify or damp the cycle. If $\alpha=0$, then $u=A$ and the oscillation of v is everywhere amplified in the ratio $A:1$. If $\alpha<0$, then $u=Ae^{\alpha x}$ decreases exponentially to zero as x increases, and the oscillations of v are diminished as x increases; this is the case of damping. If $\alpha>0$, then $u=Ae^{\alpha x}$ increases exponentially as x increases and the amplification of v increases to match; this is the anti-damping case. The oscillatory term $v=\cos(\omega x-\epsilon)$ is represented graphically in Fig. 14.4*b*; it is the graph of the function $\cos x$ with the x variable re-scaled and measured from another origin. A peak $\cos x=1$ occurs where $x=0$; a peak $\cos(\omega x-\epsilon)=1$ occurs where $x=\epsilon/\omega$. Hence the *phasing* of $v=\cos(\omega x-\epsilon)$ is fixed by the peak $v=1$ at $x=\epsilon/\omega$. The re-scaling of the x variable is such that, whereas $\cos x$ completes a cycle in the interval $0\leqslant x\leqslant 2\pi$ and then repeats in every interval of 2π, $\cos(\omega x-\epsilon)$ has a repeating cycle over an interval of $2\pi/\omega$. This is the *period* of $v=\cos(\omega x-\epsilon)$. The phase and period are indicated graphically in Fig. 14.4*b*.

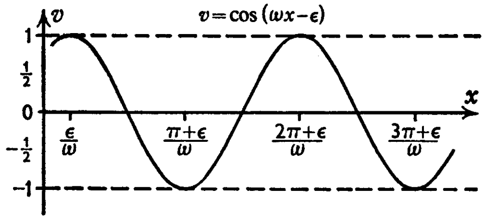


FIG. 14.4*b*

We now combine the two terms of $y=uv=Ae^{\alpha x}\cos(\omega x-\epsilon)$. If $\alpha=0$, the cosine cycle of Fig. 14.4*b* is amplified by the *amplitude* A; its shape is unchanged except that it oscillates between ± A instead of ± 1. If $\alpha < 0$, the cosine cycle is progressively diminished in amplitude as x increases, according to the *damping* factor $(-\alpha)$, the case illustrated in Fig. 14.4*c*. Similarly, if $\alpha > 0$, the cosine cycle is

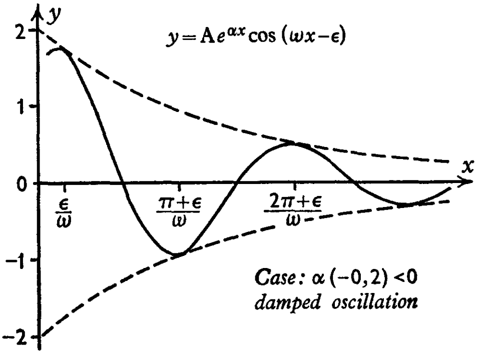


Fig. 14.4c

progressively amplified, according to the anti-damping factor α. The variation of y is similar to that of Fig. 14.4*c*, except that the oscillation is not damped but rather explosive (or anti-damped) as x increases. Hence:

Theorem: *The circular or sinusoidal function* $y = Ae^{\alpha x}\cos(\omega x - \epsilon)$ *has a symmetric oscillation of the cosine form with the features:* **Period** *T given by $2\pi/\omega$;* **Phase** *given by a peak at $x = \epsilon/\omega$;* **Amplitude** *given initially by A;* **Damping** *indicated by $(-\alpha)$.*

The period $2\pi/\omega$ represents the interval of x over which a complete cycle of y takes place; the cycle is then repeated in each successive interval of length $2\pi/\omega$. An alternative expression of the same feature is by specification of the 'frequency' of the oscillation, i.e. the number of times the cycle repeats in a unit interval of x. The frequency is $\omega/2\pi$ cycles per unit of x, or ω cycles per interval 2π of x. Hence, we can speak of the *frequency* ω of $y = Ae^{\alpha x}\cos(\omega x - \epsilon)$ as compared with the unit frequency of the cosine function $\cos x$. The frequency is the number of complete cycles achieved by the function in an interval 2π of x. The frequency and period are reciprocal to each other; it is useful to use period when the cycle is long and frequency when it is short.

To summarise: the linear differential equation

$$D^2y - 2\alpha Dy + (\alpha^2 + \omega^2)y = 0,$$

where α and ω are structural constants, has the oscillatory solution $y = Ae^{\alpha x}\cos(\omega x - \epsilon)$ where A and ϵ are arbitrary constants. The structure of the equation fixes the period $2\pi/\omega$ of oscillation and the extent of the damping $(-\alpha)$. The arbitrary constants (i.e. the initial conditions) fix the amplitude A and the phase (given by ϵ/ω). No matter what initial conditions are imposed, the oscillation of y given by the differential equation has a fixed period and a fixed damping.

14.5. The use of the operator D. The operator $D = \dfrac{d}{dx}$ for derivative and its inverse D^{-1} for anti-derivative or integral are extended to form a group under multiplication (10.8 above) by writing powers D^n for the nth derivative (n positive) or the $(-n)$th integral (n negative). It is perfectly possible, simply by imposing the appropriate definitions, to extend further and to incorporate sums and scalar products. The extension is most easily seen by taking it in two stages.

NOTATION: *The polynomial operator $D^n + a_1D^{n-1} + \dots + a_{n-1}D + a_n$ is such that*

$$(D^n + a_1D^{n-1} + \dots + a_{n-1}D + a_n)y = D^ny + a_1D^{n-1}y + \dots + a_{n-1}Dy + a_n$$

for any positive integer n and for any function y with derivatives up to the nth.

With this notation, it is found that polynomial operators follow the same algebraic processes as algebraic polynomials. For example:

(i) $(D^2 + D)y = D^2y + Dy$ by the notation. Further:

$$D(D+1)y = D(Dy + y) = D^2y + Dy = (D^2 + D)y.$$

Hence the operator $D^2 + D$ can be factorised $D(D+1)$. Similarly:

$$\begin{aligned}(D-1)(D+1)y &= (D-1)(Dy+y) = D(Dy+y) - (Dy+y)\\ &= D^2y + Dy - Dy - y = D^2y - y = (D^2-1)y\end{aligned}$$

and so the operator $(D-1)(D+1) = D^2 - 1$ and again factorisation is valid.

The next notation completes the extension:

NOTATION: *The rational fraction operator $F(D) = \dfrac{F_1(D)}{F_2(D)}$, where $F_1(D)$ and $F_2(D)$ are polynomial operators, is such that:*

$$\textit{if}\quad \frac{F_1(D)}{F_2(D)}y=z, \quad \textit{then} \quad F_1(D)y=F_2(D)z.$$

This is, in effect, an extension of the use of $\frac{1}{D}=D^{-1}$ for anti-derivative. For, if $\frac{1}{D}y=z$, then $y=Dz$, i.e. if $\int y\,dx=z$ then $y=\frac{dz}{dx}$. To illustrate:

$$\text{(ii)}\ \frac{D^2+1y}{D}=D^{-1}(D^2y+y)=D^{-1}D^2y+D^{-1}y=Dy+D^{-1}y$$
$$=(D+D^{-1})y$$

and so $\frac{D^2+1}{D}=D+D^{-1}$, as obtained by dividing through by D. Similarly:

$$\frac{D^2-1}{D+1}y=z \quad \text{means} \quad (D^2-1)y=(D+1)z.$$

Hence, $(D+1)(D-1)y=(D+1)z$

i.e. $(D-1)y=z$ since $D+1$ applied to each side gives equivalent expressions.

So: $\frac{D^2-1}{D+1}=D-1$, as obtained by dividing through by $D+1$.

These examples serve to show that manipulations of operators which are polynomials (or ratios of polynomials) in $D=\frac{d}{dx}$ follow all the algebraic processes. Such operators are applied to any function with the appropriate number of derivatives. They are to be written before the function, and never following it, since Dy has meaning but yD not.

In technical language, the new notations achieve the adjunction of an outside element D to the field of real numbers (the a's). The result is a rational fraction operator $F(D)$ and all such operators form a field, with the appropriate sums and products. Indeed, the operators not only form a field, but also a vector space over the field of real numbers. They are like complex numbers or rational (algebraic) fractions, forming a set which is a field with a scalar product operation, i.e. a vector space with the structure of a field. In short, they are perfectly well-behaved algebraically.

It is useful in practice to have the result of applying a rational fraction operator in D to various specific functions. Two simple cases are given here, for the exponential and circular functions. Other cases can be given, as in 14.9 Ex. 7.

Exponential functions. From the standard form, $De^{\lambda x}=\lambda e^{\lambda x}$, we get:

$$F(D)e^{\lambda x}=F(\lambda)e^{\lambda x} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Here $F(x)$ is a rational fraction, $F(\lambda)$ is the real value obtained by substituting λ for x and $F(D)$ is the corresponding operator in D. It is assumed in (1) that the denominator in $F(\lambda)$ is not zero. The result (1) can be summarised by saying that, in writing any expression including derivatives of an exponential function $e^{\lambda x}$, we can substitute $D=\lambda$. See 14.9 ex. 6.

Circular functions. From the standard forms, $D\cos\omega x=-\omega\sin\omega x$ and $D\sin\omega x=\omega\cos\omega x$, we get:

$$D^2\cos\omega x=-\omega^2\cos\omega x \quad \text{and} \quad D^2\sin\omega x=-\omega^2\sin\omega x.$$

So: and

$$\left.\begin{aligned} F(D^2)\cos\omega x&=F(-\omega^2)\cos\omega x\\ F(D^2)\sin\omega x&=F(-\omega^2)\sin\omega x\end{aligned}\right\}\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

In (2), the polynomial or rational fraction $F(D^2)$ consists only of even powers of D. The result can be summarised by saying that, for derivatives of a circular function $\cos\omega x$ or $\sin\omega x$, we can substitute $D^2=-\omega^2$.

Some examples illustrate (1) and (2) in practice:

(iii) $(D^2-1)e^{\lambda x}=(\lambda^2-1)e^{\lambda x}$, which is checked:

$$(D^2-1)e^{\lambda x}=D^2e^{\lambda x}-e^{\lambda x}=\lambda^2e^{\lambda x}-e^{\lambda x}=(\lambda^2-1)e^{\lambda x}.$$

Further, $\dfrac{1}{D^2-1}e^{\lambda x}=\dfrac{e^{\lambda x}}{\lambda^2-1}$. In checking, if $\dfrac{1}{D^2-1}e^{\lambda x}=z$, then the notation means that $(D^2-1)z=e^{\lambda x}$. But

$$\left(D^2-1\right)\left(\frac{e^{\lambda x}}{\lambda^2-1}\right)=\frac{1}{\lambda^2-1}\left(D^2-1\right)e^{\lambda x}=\frac{1}{\lambda^2-1}\left(\lambda^2-1\right)e^{\lambda x}=e^{\lambda x}.$$

Hence, $z=\dfrac{e^{\lambda x}}{\lambda^2-1}$ as required.

(iv) $(D^2+1)\cos\omega x=(1-\omega^2)\cos\omega x$, which is obtained directly:

$$\begin{aligned}(D^2+1)\cos\omega x&=D^2\cos\omega x+\cos\omega x\\ &=-\omega^2\cos\omega x+\cos\omega x=(1-\omega^2)\cos\omega x.\end{aligned}$$

Again, $\frac{1}{D^2+1}\sin\omega x=\frac{\sin\omega x}{1-\omega^2}$, to be checked by showing that

$$\left(D^2+1\right)\left(\frac{\sin\omega x}{1-\omega^2}\right)=\sin\omega x$$

and this is so.

The development of the operator D in this way is designed to lead to a practical method of solving linear differential equations with constant coefficients. Write the homogeneous form of an nth order equation of this type:

$$F(D)y=0 \quad \text{where } F(D)=D^n+a_1D^{n-1}+\ldots+a_{n-1}D+a_n \ldots\ldots(3)$$

Factorise the polynomial $F(D)$ into:

$$F(D)=(D-\lambda_1)(D-\lambda_2)\ldots(D-\lambda_n)$$

where $\lambda_1, \lambda_2, \ldots \lambda_n$ are the roots (real or complex) of the auxiliary equation $F(D)=0$. This means no more than that the auxiliary equation $F(\lambda)=0$ has these n roots. Then, if $D=\lambda_1$, $F(D)=0$ and we have a particular solution of (3). For $D=\lambda_1$ implies that a solution y_1 is such that $Dy_1=\lambda_1y_1$, i.e. such that $y_1=e^{\lambda_1x}$. Similarly for $\lambda_2, \lambda_3, \ldots \lambda_n$. Hence we obtain the result of 14.3, generalised and in a slightly different form:

THEOREM: *The general solution of $F(D)y=0$ where*

$$F(D)=D^n+a_1D^{n-1}+\ldots+a_{n-1}D+a_n$$

is:

$$y=A_1e^{\lambda_1x}+A_2e^{\lambda_2x}+\ldots+A_ne^{\lambda_nx}$$

where $D=\lambda_1, \lambda_2, \ldots \lambda_n$ are the roots of $F(D)=0$, provided they are all different, and where $A_1, A_2, \ldots A_n$ are arbitrary constants.

The case of multiple roots of $F(D)=0$ raises the difficulty met and solved in 14.3. If $D=\lambda$ is a double root of $F(D)=0$, then the corresponding part of the solution y is $(A_1+A_2x)e^{\lambda x}$. In this way, the correct number of arbitrary constants is maintained. The result generalises to triple and higher multiple roots (14.9 Ex. 8).

Next, write the non-homogeneous equation $F(D)y=\phi(x)$, where $F(D)$ is the same polynomial operator as before. We can also write:

$$y=\frac{1}{F(D)}\phi(x) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

since this notation simply means that $F(D)y=\phi(x)$. However, if $\phi(x)$ is of suitable form, we can apply results such as (1) and (2) above to transform (4) into a function of x, derived from $\phi(x)$. The result is a particular integral of the differential equation $F(D)y=\phi(x)$.

THEOREM: *A particular integral of $F(D)y=\phi(x)$ is to be obtained as a function of x from: $y=\frac{1}{F(D)}\phi(x)$ according to the form of $\phi(x)$. The general solution of the differential equation is then the particular integral added to the complementary function of the previous theorem.*

The application of this theorem depends entirely on having operator results of the kind shown in (1) and (2). The following examples illustrate:

(v) $D^2y-y=e^{2x}$. The complementary function is obtained from the roots $D=\pm 1$ of the auxiliary equation $D^2-1=0$; it is

$$y=A_1e^x+A_2e^{-x}.$$

The particular integral is given by

$$y=\frac{1}{D^2-1}e^{2x}=\frac{e^{2x}}{2^2-1}=\tfrac{1}{3}e^{2x}.$$

The complete solution is: $y=A_1e^x+A_2e^{-x}+\frac{1}{3}e^{2x}$.

(vi) $D^2y-y=\cos\omega x$. The complementary function is that of example (v). The particular integral is $y=\frac{1}{D^2-1}\cos\omega x=\frac{\cos\omega x}{-\omega^2-1}$ and the complete solution is: $y=A_1e^x+A_2e^{-x}-\frac{\cos\omega x}{1+\omega^2}$.

14.6. Linear difference equations. A problem may be put in terms of the rate of growth of a function $y=f(x)$ as x increases continuously. Alternatively, it may be expressed as the change in the variable y over a regular sequence of discrete values of x or over discrete intervals of x. If the unit for x is selected as the fixed interval, we can write $y_n=f(n)$ for $n=0, 1, 2, 3, \ldots$ from a starting point y_0 at $n=0$. A case in point is the familiar problem of compound interest when interest is compounded annually at $100a$ per cent per year. The result for the amount £y_n of an initial £y_0 after n years is:

$$y_n=y_0(1+a)^n.$$

This is to be derived from the fact that, in one year after the nth, £y_n grows to £y_{n+1} by the addition of £ay_n of interest: $y_{n+1}-y_n=ay_n$. Hence, write:

$$y_{n+1}-(1+a)y_n=0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

as a 'difference equation' from which to derive $y_n = y_0(1+a)^n$ in terms of an arbitrary constant or initial value y_0. This matches the differential equation of 14.2 in the corresponding problem of interest compounded continuously.

The theory of difference equations, of which (1) is a simple example, follows very closely the corresponding theory of differential equations.

DEFINITION: *An (ordinary)* **difference equation** *is a relation between successive values of a discrete variable y_n for $n = 0, 1, 2, \ldots$:*

$$F(x, y_n, y_{n+1}, \cdots y_{n+r}) = 0$$

where r is the **order** *of the equation. The equation is linear if it is:*

$$y_{n+r} + a_1 y_{n+r-1} + \ldots + a_{r-1} y_{n+1} + a_r y_n = \phi(n)$$

where the coefficients are constants or dependent on n.

Following the argument of 14.2, we expect that any difference equation of order r has a solution with r arbitrary constants, each of which can be eliminated by one process of differencing. Further, a linear equation

$$y_{n+r} + a_1 y_{n+r-1} + \ldots + a_{r-1} y_{n+1} + a_r y_n = \phi(n) \quad \ldots\ldots\ldots\ldots(2)$$

has a corresponding *homogeneous form:*

$$y_{n+r} + a_1 y_{n+r-1} + \ldots + a_{r-1} y_{n+1} + a_r y_n = 0 \quad \ldots\ldots\ldots\ldots\ldots(3)$$

It then follows that, if $y_n = f_1(n)$ and $y_n = f_1(n)$ are two solutions of (3), so is $y_n = A_1 f_1(n) + A_2 f_2(n)$ for any constants A_1 and A_2. Further, if $y_n = \bar{f}(n)$ is a particular solution of (2) and $y_n = f_1(n)$ a solution of (3), then $y_n = f_1(n) + \bar{f}(n)$ is also a solution of (2). Hence:

THEOREM: *The general solution of the linear difference equation* (2) *is:*

$$y_n = A_1 f_1(n) + A_2 f_2(n) + \ldots + A_r f_r(n) + \bar{f}(n) \ldots\ldots\ldots\ldots\ldots(4)$$

where $f_1(n), f_2(n), \ldots f_r(n)$ making up the **complementary function** *are r different solutions of the homogeneous form* (3), *where $\bar{f}(n)$ is the* **particular integral,** *any solution of the equation* (2), *and where A_1, $A_2, \ldots A_r$ are arbitrary constants.*

The arbitrary constants can be expressed in terms of r initial conditions, usually expressed as the given initial values of y_n:

$$y_0 \text{ at } n=0;\ y_1 \text{ at } n=1;\ \ldots\ y_{r-1} \text{ at } n=r-1.$$

With these initial values, the difference equation (2) provides, by a process of iteration, each succeeding value: $y_r, y_{r+1}, y_{r+2}, \ldots$. For (2) gives y_r in terms of $y_0, y_1, \ldots y_{r-1}$ (given), then y_{r+1} in terms of $y_1, y_2, \ldots y_r$, and so on. This is always possible, but it often fails to give y_n explicitly as a function of n. An example illustrates:

(i) $y_{n+1} - 2y_n = n + 2$, given $y_0 = 2$ at $n = 0$.

In succession: $y_1 = 2y_0 + 2 = 6$; $y_2 = 2y_1 + 3 = 15$; $y_3 = 2y_2 + 4 = 34$; ... Hence, for $n = 0, 1, 2, 3, \ldots, y_n = 2, 6, 15, 34, \ldots$. This still does not make it clear what y_n is as a function of n. Other methods need to be sought for this.

When the linear difference equation has *constant coefficients*, a general method can be devised for writing the solution (4). The *first-order* equation is: $y_{n+1} + ay_n = \phi(n)$. The particular integral is any one solution which can be found. The complementary function is to be got from the homogeneous form: $y_{n+1} = (-a)y_n$. Hence

$$y_n = (-a)y_{n-1} = (-a)^2 y_{n-2} = \ldots,$$

giving:
$$y_n = A(-a)^n \quad n = 0, 1, 2, \ldots$$

where A is arbitrary. This is to be compared with the solution $y = Ae^{-ax}$ for the corresponding differential equation $Dy + ay = 0$. The power function $(-a)^x$, with x integral, appears instead of the exponential e^{-ax}. The form of the solution depends on the value of the constant a. The variable y_n increases or decreases in absolute value according as $|a| > 1$ or $|a| < 1$; the sign of a then determines whether y_n varies steadily or by alternation in sign. To return to the example:

(ii) $y_{n+1} - 2y_n = n + 2$ has complementary function $y_n = A2^n$. For a particular integral, try $y_n = \alpha n + \beta$ and attempt to find α and β. So:

$$y_{n+1} - 2y_n = \alpha(n+1) + \beta - 2(\alpha n + \beta) = -\alpha n + (\alpha - \beta).$$

This is to be equal to $n + 2$ for all n, i.e. $-\alpha = 1$ and $\alpha - \beta = 2$. Hence, $\alpha = -1$, $\beta = -3$. The particular integral is $y_n = -(n+3)$. The complete solution is:

$$y_n = A2^n - (n+3).$$

Given $y_n = y_0$ at $n = 0$: $y_0 = A - 3$. Hence:

$$y_n = (y_0 + 3)2^n - (n+3).$$

If $y_0 = 2$, then $y_n = 5 \,.\, 2^n - (n+3)$ for $n = 0, 1, 2, \ldots$
i.e. the sequence $y_n = 2, 6, 15, 34, \ldots$ of example (i) above.

The *second-order* difference equation $y_{n+1}+ay_{n+1}+by_n=\phi(n)$ has the homogeneous form:

$$y_{n+2}+ay_{n+1}+by_n=0 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

Apart from a particular integral (to be obtained however we can), the solution of the equation reduces to getting the complementary function from (5). The solution in the first-order case suggests trying $y_n=\lambda^n$ for appropriate λ. From (5):

$$\lambda^{n+2}+a\lambda^{n+1}+b\lambda^n=0.$$

If $\lambda=0$, then nothing is added to the particular integral. Hence, take $\lambda\neq 0$ and cancel λ^n to get the *auxiliary equation:* $\lambda^2+a\lambda+b=0$ with two roots λ_1 and λ_2. The complementary function is

$$y_n=A_1\lambda_1{}^n+A_2\lambda_2{}^n \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(6)$$

The form of (6) depends on whether λ_1 and λ_2 are real or complex.

Case: $a^2>4b$. Real and distinct roots: $\lambda_1,\lambda_2=\frac{1}{2}\{-a\pm\sqrt{(a^2-4b)}\}$.

The solution of (5) is then (6) with real λ_1 and λ_2 as specified. It is very similar to the solution of the first-order case, with two power function terms to consider instead of one.

Case: $a^2=4b$. Real and equal roots: $\lambda=-\frac{1}{2}a=-\sqrt{b}$.

The equation (5) is then: $y_{n+2}-2\lambda y_{n+1}+\lambda^2 y_n=0$. Only one particular solution ($y_n=\lambda^n$) is found so far. To get another, adopt the trick which worked in 14.2 and try $y_n=n\lambda^n$:

$$\begin{aligned} y_{n+2}-2\lambda y_{n+1}+\lambda^2 y_n &= (n+2)\lambda^{n+2}-2\lambda(n+1)\lambda^{n+1}+\lambda^2 n\lambda^n \\ &= \{(n+2)-2(n+1)+n\}\lambda^{n+2}=0 \end{aligned}$$

i.e. $y_n=n\lambda^n$ is the second solution and the general solution of (5) is:

$$y_n=(A_1+nA_2)\lambda^n.$$

Again, this is not very different from the solution of a first-order equation. The sign and magnitude of λ determines the form of y_n.

Case: $a^2<4b$. Conjugate complex roots: $\lambda_1,\lambda_2=\frac{1}{2}\{-a\pm i\sqrt{(4b-a^2)}\}$.

Write the roots in the form: $\lambda_1,\lambda_2=r(\cos\theta\pm i\sin\theta)=re^{\pm i\theta}$

where $\qquad r\cos\theta=-\frac{1}{2}a \quad$ and $\quad r\sin\theta=\frac{1}{2}\sqrt{(4b-a^2)}$

i.e. where $\qquad r=\sqrt{b} \quad$ and $\quad \tan\theta=-\sqrt{\dfrac{4b}{a^2}-1}$

given in terms of the structural constants a and b of the equation (5). The solution of (5) is then expressed:

$$y_n = A_1\lambda_1{}^n + A_2\lambda_2{}^n = A_1 r^n e^{in\theta} + A_2 r^n e^{-in\theta}$$
$$= r^n\{A_1(\cos n\theta + i \sin n\theta) + A_2(\cos n\theta - i \sin n\theta)\}$$
$$= r^n(B_1 \cos n\theta + B_2 \sin n\theta)$$

where $B_1 = A_1 + A_2$ and $B_2 = i(A_1 - A_2)$ are alternative arbitrary constants. A further switch in arbitrary constants to A and ϵ is made:

$$B_1 = A \cos \epsilon \quad \text{and} \quad B_2 = A \sin \epsilon$$

i.e. $$A = \surd(B_1^2 + B_2^2) \quad \text{and} \quad \epsilon = \tan^{-1}(B_2/B_1).$$

So: $$y_n = Ar^n(\cos n\theta \cos \epsilon + \sin n\theta \sin \epsilon)$$

i.e. $$y_n = Ar^n \cos (n\theta - \epsilon) \text{ ....(7)}$$

This is the general form of the solution of (5) in this case. It is to be compared with the solution $y = Ae^{ax} \cos (\omega x - \epsilon)$ of the corresponding differential equation in 14.4. Again (7) represents an oscillatory variation in y_n for $n = 0, 1, 2, \ldots$ and the variation is of sinusoidal form. The *period* is $2\pi/\theta$ where θ is given by $\tan \theta = -\sqrt{\frac{4b}{a^2} - 1}$ in terms of structural constants. The *damping* is described by the positive constant $r = \surd b$, again given by the structure of the equation. The oscillation is damped if $r < 1$, regular if $r = 1$ and explosive (anti-damped) if $r > 1$. The *amplitude* A and *phasing* ϵ are arbitrary constants, given by initial conditions. An example illustrates:

(iii) $y_{n+2} - y_{n+1} + \frac{1}{2}y_n = (\frac{1}{2})^n$.

The complementaty function is obtained from the auxiliary equation $\lambda^2 - \lambda + \frac{1}{2} = 0$ with roots $\frac{1}{2}(1 \pm i)$. Hence:

$$r \cos \theta = \tfrac{1}{2} \quad \text{and} \quad r \sin \theta = \tfrac{1}{2} \quad \text{i.e. } r = \frac{1}{\surd 2} \text{ and } \tan \theta = 1.$$

Explicitly, $\theta = \frac{\pi}{4}$, the value of θ in the range 0 to 2π for $\tan \theta = 1$ (and $\sin \theta = \cos \theta = \frac{1}{\surd 2}\Big)$. The complementary function is:

$$y_n = A\left(\frac{1}{\surd 2}\right)^n \cos \left(\frac{n\pi}{4} - \epsilon\right)$$

for arbitrary A and ϵ, to be given by initial conditions. This is a damped oscillation with period $2\pi/\theta$, where $\theta = \pi/4$. The period is

8 units, i.e. the cycle is complete in the range from $n=0$ to $n=8$. To find a particular integral, try $y_n=k(\frac{1}{2})^n$ for some k. Substitute:

$$y_{n+2}-y_{n-1}+\tfrac{1}{2}y_n=\tfrac{1}{4}k(\tfrac{1}{2})^n \quad \text{to equal } (\tfrac{1}{2})^n \text{ for all } n.$$

Hence $k=4$. The complete solution of the equation is:

$$y_n=4(\tfrac{1}{2})^n+A\left(\frac{1}{\sqrt{2}}\right)^n\cos\left(\frac{n\pi}{4}-\epsilon\right).$$

As n increases, the first (steady) term dies out, as does the damped oscillation of period 8.

Linear difference equations with constant coefficients and of order higher than the second are solved in the same way. It is necessary to find the roots (more than two in number) of an auxiliary equation of higher order than a quadratic. The complementary function in the solution then contains several terms, some of which may be of the form (6) with real λ's and others of the oscillatory form (7). Just as the operator D is of use in solving differential equations (14.5), so now an operator can be introduced in the solution of difference equations. It is the *shift operator* E defined so that $Ey_n=y_{n+1}$, i.e. E is the operation of getting from y_n to the next value y_{n+1} in sequence. This is examined in 14.9 Ex. 17.

Simultaneous linear difference equations in several variables can be handled on the same lines as for differential equations. A simple case is the pair of linear difference equations in two discrete variables y_n and z_n:

$$y_{n+1}=a_{11}y_n+a_{12}z_n \quad \text{and} \quad z_{n+1}=a_{21}y_n+a_{22}z_n \quad \dots\dots\dots\dots(8)$$

These are such that y_n and z_n each satisfies the same difference equation of the second order. Both variables follow the same type of path; for example, both may have an oscillation of the same period and damping. See 14.9 Ex. 13.

14.7. Laplace Transforms. The linear (algebraic) transformation has the additive property that the transform of a linear combination is the linear combination of the separate transforms (14.1 above). The same property holds for the derivative $f'(x)$ of a function $f(x)$, taking this as a transform, by the operator D, of one function into another. If $f_1(x)$ has derivative $f_1'(x)$ and $f_2(x)$ derivative $f_2'(x)$, then $\lambda_1 f_1(x)+\lambda_2 f_2(x)$ has derivative $\lambda_1 f_1'(x)+\lambda_2 f_2'(x)$ for any constants λ_1

and λ_2. Other transforms from one function to another have this basic additive property; one of them is now considered. It is the Laplace Transform named after Laplace (1749–1827) and it has many practical uses.

DEFINITION: *The* **Laplace Transform** *of the function $y(x)$ defined for $x \geqslant 0$ is:*

$$\bar{y}(p) = \int_0^\infty e^{-px} y(x)\, dx = \underset{k\to\infty}{Lim} \int_0^k e^{-px} y(x)\, dx$$

defined for those real values of p for which the infinite integral exists.

The choice of notation for the new function of p, derived as the Laplace Transform of a given function $y(x)$, is not an easy one to make.* The notation, adopted here, is in quite general use and it does serve to stress, by putting a 'bar' over y, that y is transformed into a new function $\bar{y}$.

Given a function $y(x)$, or a class of functions, it is necessary first to determine whether the infinite integral for $\bar{y}(p)$ exists. It may exist for some functions and not for others, or for some values of p and not for others. Note that $e^{-px} \to 0$ as $x \to \infty$ so rapidly when p is positive (12.3 above) that it usually outweighs any tendency for $y(x)$ to increase with x. Hence, the Laplace Transform is to be expected to exist for most ordinary functions if p is positive (and sufficiently large). The transform fails only when p is small (or negative) and/or for such unusual functions as $y(x) = e^{x^2}$. Some examples illustrate how $\bar{y}(p)$ is to be obtained from the definition:

(i) If $y(x) = 1$ (constant), then $\bar{y}(p) = \int_0^\infty e^{-px}\, dx = \left[-\frac{1}{p} e^{-px}\right]_0^\infty = \frac{1}{p}$ since $e^{-px} \to 0$ as $x \to \infty$ $(p > 0)$. So $y(x) = 1$ has Laplace Transform $\bar{y}(p) = \frac{1}{p}$ $(p > 0)$.

(ii) If $y(x) = x$, then $\bar{y}(p) = \int_0^\infty x e^{-px}\, dx$. Integrate by parts:

$$\int x e^{-px}\, dx = x \int e^{-px}\, dx - \int \left(Dx \int e^{-px}\, dx \right) dx = -\frac{1}{p} x e^{-px} + \frac{1}{p} \int e^{-px}\, dx$$

* See J. C. Jaeger: *An Introduction to the Laplace Transformation* (Methuen, 1949), p. vi.

i.e.
$$\int xe^{-px}\,dx = -\frac{1}{p}e^{-px}\left(x+\frac{1}{p}\right).$$

So: $\displaystyle\int_0^\infty xe^{-px}\,dx = \frac{1}{p^2}$ since $e^{-px}\to 0$ and $xe^{-px}\to 0$ as $x\to\infty$ $(p>0)$.

The Laplace Transform of $y(x)=x$ is $\bar{y}(p)=\dfrac{1}{p^2}$ $(p>0)$.

(iii) If $y(x)=x^n$, then $\bar{y}(p)=\displaystyle\int_0^\infty x^n e^{-px}\,dx = I_n$ (say). Now:

$$\int x^n e^{-px}\,dx = x^n\int e^{-px}\,dx - \int\left(Dx^n\right)\left(\int e^{-px}\,dx\right)dx =$$
$$-\frac{1}{p}x^n e^{-px} + \frac{n}{p}\int x^{n-1}e^{-px}\,dx.$$

Since $x^n e^{-px}\to 0$ as $x\to\infty$, insertion of the bounds of integration gives:

$$I_n = \frac{n}{p}I_{n-1} \quad (p>0).$$

So: $\displaystyle I_n = \frac{n}{p}\,\frac{n-1}{p}I_{n-2} = \frac{n}{p}\,\frac{n-1}{p}\,\frac{n-2}{p}I_{n-3} = \ldots = \frac{n!}{p^n}I_0 = \frac{n!}{p^{n+1}}.$

Since $I_0=\displaystyle\int_0^\infty e^{-px}\,dx = \frac{1}{p}$ as in (i) above. Hence, the Laplace Transform of $y(x)=x^n$ is $\bar{y}(p)=\dfrac{n!}{p^{n+1}}$ $(p>0)$.

The basic additive property of Laplace Transforms follows from the definition:

THEOREM: *If $y_1(x)$ and $y_2(x)$ have Laplace Transforms $\bar{y}_1(p)$ and $\bar{y}_2(p)$, then $\lambda_1 y_1(x)+\lambda_2 y_2(x)$ has Laplace Transform $\lambda_1\bar{y}_1(p)+\lambda_2\bar{y}_2(p)$ for any constants λ_1 and λ_2.*

The proof is immediate, depending only on the additive property of integrals.

Another consequence of the definition is that, if $\bar{y}(p)$ exists for a given $y(x)$, then it is unique.* The practical problem is to find it, to write it down as quickly as possible. This is precisely the problem of the handling of derivatives in practice. First, it is to be checked that a given function $y(x)$ has a Laplace Transform, a similar problem

* The converse is also true though not proved here. Given a function $\bar{y}(p)$, then a unique $y(x)$ exists with $\bar{y}(p)$ as its Laplace Transform.

(involving limits) to that of determining whether a derivative exists. Second, a set of standard forms for the Laplace Transforms of the simplest functions is to be obtained from the definition. Third, certain operational rules are established, again from the definition, with the object of facilitating the writing of Laplace Transforms of more complicated functions.

The *standard forms* include the following:

$y(x)$	*Laplace Transform* $\bar{y}(p)$		$y(x)$	*Laplace Transform* $\bar{y}(p)$	
x^n	$\dfrac{n!}{p^{n+1}}$	$(p>0)$	$\sin \alpha x$	$\dfrac{\alpha}{p^2+\alpha^2}$	$(p>0)$
e^{ax}	$\dfrac{1}{p-a}$	$(p>a)$	$\cos \alpha x$	$\dfrac{p}{p^2+\alpha^2}$	$(p>0)$

Here a and α are constants. The proof of the first of these forms is given in example (iii) above. As particular cases ($n=0$, 1), note that 1 has Laplace Transform $1/p$ and that x has Laplace Transform $1/p^2$. The other standard forms follow from the relevant integrals, see 14.9 Exs. 19 and 20.

The first of the *operational rules* is the basic additive property. Extended to several terms, the rule is that $\lambda_1 y_1(x)+\lambda_2 y_2(x)+\ldots+\lambda_n y_n(x)$ has Laplace Transform $\lambda_1\bar{y}_1(p)+\lambda_2\bar{y}_2(p)+\ldots+\lambda_n\bar{y}_n(p)$, where λ_r is any constant and $\bar{y}_r(p)$ is the Laplace Transform of $y_r(x)$ for $r=1$, 2, ... n. For example, the Laplace Transform of $(1+x)^2=1+2x+x^2$ is $\dfrac{1}{p}+\dfrac{2}{p^2}+\dfrac{2}{p^3}$. Another operational rule easily established (14.9 Ex. 21) is that, if $y(x)$ has Laplace Transform $\bar{y}(p)$ and if a is any constant, then $e^{ax}y(x)$ has Laplace Transform $\bar{y}(p-a)$. This serves to extend the standard forms to give $\dfrac{n!}{(p-a)^{n+1}}$ as the Laplace Transform of x^ne^{ax}, $\dfrac{\alpha}{(p-a)^2+\alpha^2}$ as that of $e^{ax}\sin \alpha x$ and $\dfrac{p-a}{(p-a)^2+\alpha^2}$ as that of $e^{ax}\cos \alpha x$ (all for $p>a$).

The Laplace Transform is designed to handle functions of a continuous variable, and particularly exponential and sinusoidal functions of the kind met in solving differential equations. From the standard forms above, it is evident that the transform serves to replace such functions by algebraic expressions (ratios of polynomials) in p. The Laplace Transform, in one of its main applications, reduces a differential equation to an algebraic equation; it is often the best way of solving differential equations when initial conditions are specified. There is a corresponding transform designed to handle a discrete variable, and to solve difference equations. It is the *Generating Function* which transforms a given sequence y_n $(n=0, 1, 2, \dots)$ into a function of s:

$$\bar{y}(s)=\sum_{n=0}^{\infty} y_n s^n.$$

As compared with the Laplace Transform, this has the power function s^n instead of the exponential e^{-px} and an infinite series instead of an infinite integral. It is also a transform with the basic additive property: if the sequence y_n has transform $\bar{y}(s)$ and if the sequence z_n has transform $\bar{z}(s)$, then the sequence $(\lambda_1 y_n+\lambda_2 z_n)$ has transform $\lambda_1\bar{y}(s)+\lambda_2\bar{z}(s)$ for any constants λ_1 and λ_2.

14.8. Linear models. In many problems the variables are related among themselves by means of inter-dependent equations, usually differential or difference equations. A frequent case is that in which the variables are functions of time such that one variable is dependent on the sequence of past values of another variable. This is the case of a *lagged dependence*. Suppose that there are two variables in a system, $y(t)$ and $z(t)$ as functions of time t. Suppose further that $y(t)$ depends on all past values of $z(t)$, the dependence being *linear* in that the influences of each past time sum in their effect on $y(t)$ after multiplication by an appropriate factor. These factors, for various past times, make up a *weighting function* $w(\tau)$, where τ is the time interval taken in the past. Hence, the current value $y(t)$ is the sum of the past values $z(t-\tau)$, each weighted with $w(\tau)$, for all $\tau>0$. In limiting or integral form:

$$y(t)=\int_0^{\infty} w(\tau)z(t-\tau)\,d\tau \qquad \dots\dots\dots\dots\dots\dots\dots(1)$$

Similarly, $z(t)$ may have a lagged dependence on $y(t)$:

$$z(t)=\int_0^{\infty} w'(\tau)y(t-\tau)\,d\tau \qquad \dots\dots\dots\dots\dots\dots\dots(2)$$

where $w'(\tau)$ is another weighting function. The equations (1) and (2) together make up a linear and inter-dependent system in two variables. It is required to solve the equations to give the time-paths of $y(t)$ and $z(t)$.

Such problems arise in many fields, e.g. to describe the behaviour of physical equipment, as in control systems in engineering, or of organisms in ecological problems. Equally, they may express the behaviour of units such as groups of firms or consumers in economics. Two examples illustrate.

(i) The speed of a steam turbine is regulated by a governor which varies the amount of the steam valve opening. Because of inertia in the turbine, the speed $y(t)$ depends on the valve opening $z(t-\tau)$ at earlier times $(\tau>0)$. If the dependence is linear, then it appears as (1) for some weighting function $w(\tau)$ determined by the properties of the turbine. Further, the governor of the turbine is designed to change the valve opening according to the discrepancy between the turbine's speed and some desired speed, again with lags in the operation of the governor. The valve opening $z(t)$ depends on the turbine's speed $y(t-\tau)$ at earlier times $(\tau>0)$. Then (2) expresses such a dependence, taken as linear, where the weighting function $w'(\tau)$ is given by the design of the governor. The problem is to solve (1) and (2) to get the time-path $y(t)$, the speed of the turbine, and to vary the design of the system to eliminate or minimise fluctuations in $y(t)$.

(ii) Consider an economic problem of production and consumption in the whole economy, ignoring (for simplicity here) the matter of investment. Let $y(t)$ be the supply and $z(t)$ the demand for goods and services, both in money values. Because of lags in the production system, supply $y(t)$ depends on demand $z(t-\tau)$ at earlier times $(\tau>0)$, as shown by (1) in a linear model. Further, the money value $y(t)$ is also the aggregate income in the economy, out of which the demand $z(t)$ arises. Because of lags in the distribution and disposal of incomes, demand $z(t)$ depends on income $y(t-\tau)$ at earlier times $(\tau>0)$, given by (2) in a linear model. The problem is to solve (1) and (2) for the time-path of $y(t)$, production and income in the economy, and to stabilise the economy by modifying the lag (weighting) functions $w'(\tau)$ and $w(\tau)$ or otherwise.

A system which includes pairs of relations such as (1) and (2) is characterised by a *feed-back;* one variable is influenced by another and, in its turn, feeds back to regulate the other. In case (i), the speed of the turbine depends on the valve opening but also feeds back to regulate the valve opening. Similarly, in (ii), the simple mutual relations between supply and demand (production and consumption) constitute a feed-back system. Now consider a particular case, that of an exponential lag, or weighting function, in the dependence of one variable on another. This is not at all unrealistic:

(iii) Take $w(t)=\lambda e^{-\lambda t}$ (λ constant) in (1). The constant λ is designed to make the sum (integral) of the weights equal to unity:

$$\int_0^\infty w(t)\,dt=\lambda\int_0^\infty e^{-\lambda t}\,dt=\Big[-e^{-\lambda t}\Big]_0^\infty=1.$$

Substitute in (1) and change the variable of integration by writing $x=t-\tau$ ($dx=-d\tau$):

$$y(t)=\lambda\int_0^\infty e^{-\lambda\tau}z(t-\tau)\;d\tau=-\lambda\int_t^{-\infty}e^{-\lambda(t-x)}z(x)\,dx$$

$$=\lambda e^{-\lambda t}\int_{-\infty}^t e^{\lambda x}z(x)\,dx$$

i.e.
$$\frac{1}{\lambda}\,e^{\lambda t}y(t)=\int_{-\infty}^t e^{\lambda x}z(x)\,dx.$$

Write and equate the derivatives of the two sides:

$$e^{\lambda t}y(t)+\frac{1}{\lambda}\,e^{\lambda t}Dy(t)=e^{\lambda t}z(t)$$

i.e.
$$Dy(t)+\lambda y(t)=\lambda z(t).$$

This is a differential equation, first-order and linear, to which the given relation (1) reduces when the weighting function $w(t)$ is exponential. In operator form, the differential equation is $(D+\lambda)y=\lambda z$,

i.e.
$$y=\frac{\lambda}{D+\lambda}\,z.$$

This example suggests, what is in fact usually true, that a relation of the linear form (1) reduces to a differential equation in $y(t)$ and $z(t)$, and that the equation is linear with constant coefficients. This is the point to be pursued. Suppose that the variable $z(t)$ shows exponential growth and/or regular oscillations, so that it is of *sinusoidal form* as given in 14.4 above:

$$z(t)=Ae^{\alpha t}\cos(\omega t+\epsilon)\;\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

Generally, (3) is an oscillation of period $\frac{2\pi}{\omega}$, damped or explosive according to the sign of α; A and ϵ fix the amplitude and phasing. One particular case ($\alpha=0$, $\omega\neq 0$) gives a regular oscillation of unchanged amplitude. Another ($\alpha\neq 0$, $\omega=0$) gives a steady exponential growth or decline at the rate α: $z=z_0e^{\alpha t}$ ($z_0=A\cos\epsilon$).

Instead of (3) in real terms, it is convenient to write the complex variable:

$$Z(t)=Ae^{i\epsilon}e^{pt} \quad \text{for } p=\alpha+i\omega \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4)$$

Then: $Z(t)=Ae^{\alpha t}e^{i\epsilon}e^{i\omega t}$

$$=Ae^{\alpha t}(\cos\epsilon+i\sin\epsilon)(\cos\omega t+i\sin\omega t) \quad \text{by (4) of 12.6}$$

$$=Ae^{\alpha t}\{\cos(\omega t+\epsilon)+i\sin(\omega t+\epsilon)\} \quad \text{by (2) of 12.5.}$$

Hence, the real part of (4) is the sinusoidal variable (3), and this is what we concentrate upon. The imaginary part of (4), $Ae^{\alpha t}\sin(\omega t+\epsilon)$, is similar but usually to be ignored. We can now operate upon (4) according to the ordinary algebraic processes. First, if t is subject to a fixed delay θ:

$$Z(t-\theta)=Ae^{i\epsilon}e^{p(t-\theta)}=e^{-p\theta}Ae^{i\epsilon}e^{pt}$$

i.e.

$$Z(t-\theta)=e^{-p\theta}Z(t). \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(5)$$

Second, using the derivative operator D:

$$DZ(t)=Ae^{i\epsilon}De^{pt}=Ae^{i\epsilon}(pe^{pt})=pZ(t)$$

$$D^2Z(t)=DpZ(t)=pDZ(t)=p^2Z(t).$$

Generally:

$$D^nZ(t)=p^nZ(t) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(6)$$

The validity of this procedure can be checked (14.9 Exs. 25 and 26) by showing, not only that $z(t)$ is the real part of $Z(t)$, but that $z(t-\theta)$ is the real part of $Z(t-\theta)$ in (5) and that $z^{(n)}(t)$ is the real part of $D^nZ(t)$ in (6). Hence:

THEOREM: *The sinusoidal variable* $z(t)=Ae^{\alpha t}\cos(\omega t+\epsilon)$ *is the real part of* $Z(t)=Ae^{i\epsilon}e^{pt}$ ($p=\alpha+i\omega$), $z(t-\theta)$ *is the real part of*

$$Z(t-\theta)=e^{-p\theta}Z(t)$$

and $z^{(n)}(t)$ *is the real part of* $D^nZ(t)=p^nZ(t)$.

In terms of operators, the *delay* θ in t in $z(t)$ corresponds to multiplying $Z(t)$ by $e^{-p\theta}$, and the *nth derivative* of $z(t)$ is obtained by writing $D=p$ in $D^nZ(t)$.

These results are to be applied to the lagged dependence (1) of $y(t)$ on $z(t)$. First, as a matter of notation or definition:

DEFINITION: *The* **transfer function** *of the lagged dependence* (1) *is the Laplace Transform of the weighting function*: $\bar{w}(p)=\int_0^\infty e^{-pt}w(t)\,dt.$

The Laplace Transform is defined in 14.7 for a real variable p but the definition extends (14.9 Ex. 22) to the case where p is a complex variable. Here the transfer function $\bar{w}(p)$ is taken as a function of a complex variable p, to be identified as $p=\alpha+i\omega$, for the sinusoidal variation considered.

In (1), let $z(t)$ be a sinusoidal variable, the real part of $Z(t)=Ae^{i\epsilon}e^{pt}$ $(p=\alpha+i\omega)$ and let the corresponding $y(t)$ be the real part of $Y(t)$. Then:

$$Y(t)=\int_0^\infty w(\tau)Z(t-\tau)\,d\tau=\int_0^\infty w(\tau)e^{-p\tau}Z(t)\,d\tau \quad \text{by (5)}$$

$$=Z(t)\int_0^\infty e^{-p\tau}w(\tau)\,d\tau$$

i.e.

$$Y(t)=\bar{w}(p)Z(t) \quad \text{.............................(7)}$$

Since $D=p=\alpha+i\omega$ for sinusoidal variables, this gives:

$$Y(t)=\bar{w}(D)Z(t)$$

or taking the real parts:

$$y(t)=\bar{w}(D)z(t) \quad \text{.............................(8)}$$

This is a differential equation connecting the sinusoidal variables $y(t)$ and $z(t)$ related by (1). The original relation is reduced to a differential equation, as desired. Moreover, it is seen, from the standard forms for Laplace Transforms (14.7), that all the usual weighting functions $w(t)$ have Laplace Transforms which are ratios of polynomials in p. Hence, if it happens (as expected) that $\bar{w}(p)$ is the ratio $F(p):G(p)$ of two polynomials, then (8) is:

$$G(D)y(t)=F(D)z(t)$$

and the differential equation is linear with constant coefficients. To illustrate:

(iv) If $w(t)=\lambda e^{-\lambda t}$, then the transfer function is $\bar{w}(p)=\dfrac{\lambda}{p+\lambda}$ as given by the standard form of 14.7. Hence, by (8), the differential equation form of (1) is:

$$y(t)=\frac{\lambda}{D+\lambda}\,z(t)$$

for this particular weighting function. This agrees with the direct method of example (iii).

To return to (7), we notice that the variation assumed for $z(t)$, the real part of $Z(t)$, gives rise to a corresponding variation in $y(t)$ as the real part of $Y(t)$. It is taken that $z(t)$ is a sinusoidal variable with *given* period $2\pi/\omega$ and with *given* damping indicated by α. Hence $p=\alpha+i\omega$ is a given complex number, and $\bar{w}(p)$ is also a given complex number. Write $\bar{w}(p)=\rho e^{i\phi}$, where ρ and ϕ are the polar coordinates of the complex number. Then by (7):

$$Y(t)=\rho e^{i\phi}Ae^{i\epsilon}e^{pt}=(\rho A)e^{i(\epsilon+\phi)}e^{pt}.$$

It follows that $y(t)$, the real part of $Y(t)$, is also a sinusoidal variable, that it has the *same* period and damping as $z(t)$, given by $p=\alpha+i\omega$, and that only the amplitude and phasing are changed. Hence:

THEOREM: *A variable $y(t)$ depends on $z(t)$ by the relation* (1), *reducing to the differential equation* (8). *If $z(t)$ is a given sinusoidal variable with period $2\pi/\omega$ and damping α, then $y(t)$ is a sinusoidal variable with the same period and damping.*

The change in the amplitude and phasing depends on the transfer function $\bar{w}(p)=\rho e^{i\phi}$. The amplitude A of $z(t)$ is multiplied by ρ to give the amplitude ρA of $y(t)$. There is a shift of phase by amount ϕ (ϵ becoming $\epsilon+\phi$) in passing from $z(t)$ to $y(t)$.

To complete the linear model, take (1) and (2) together, reducing to a pair of differential equations of form (8):

$$y(t)=\bar{w}(D)z(t) \quad \text{and} \quad z(t)=\bar{w}'(D)y(t).$$

The common oscillatory movements of $y(t)$ and $z(t)$, i.e. the common period and damping, are to be found by solving simultaneous differential equations. The respective amplitudes and phasing depend on initial conditions. This is a generalisation of a particular case already considered. If the weighting functions are simple exponentials, as in examples (iii) and (iv) above, then the differential equations are both first-order. By (6) and (7) of 14.3, both $y(t)$ and $z(t)$ satisfy the same second-order differential equation, i.e. they both have the same time-path as regards period of oscillation and damping.

The characteristics of a *linear model* for variations over time can now be seen. One variable is given as a linear combination of others. Typically, a variable $y(t)$ is a linear combination of past values of another variable $z(t)$, expressed in limiting form by the integral shown in (1). The relation reduces to a differential equation, using the transfer function of the relation as in (8). This is typically linear with constant coefficients. It implies that any sinusoidal variation in $z(t)$, or any exponential growth, is reflected in a similar variation in $y(t)$. If there are two such relations in two variables, then the corresponding pair of differential equations are to be solved. When linear with constant coefficients, the solution gives the period and damping of the common variation in the two variables, i.e. the free or self-sustaining oscillation inherent in the system. Here 'oscillation' may take the particular form of an exponential growth or decline, the case $\alpha \neq 0$, $\omega = 0$. In general, it is an oscillation of a determined period and with a determined amount of damping.*

Situations of this type are not uncommon. Certainly linear models are not as restricted in scope as might be thought. Nevertheless there are serious limitations to a linear model. Differential equations which are linear with constant coefficients have solutions confined to sinusoidal variables (including exponential growth as a particular case). Hence all the variables of the model have one and the same variation, alike in period and damping. They only differ by amplitude and phase, and then because of the 'accident' of initial conditions. There is no possibility of different oscillations for different variables — or even of an oscillation of other than the symmetric 'cosine' form. There is no structural explanation of the amplitude or phasing of the oscillations. To break out of these limitations, we need to formulate the model in non-linear terms, with relations more complicated than (1) and with differential equations which are not linear with constant coefficients.

* The extension to a system of several equations in several variables is clear enough. There is also an alternative formulation, in discrete rather than continuous variables. The relation (1) is then a sum, and (8) becomes a difference equation. The solution of a system of difference equations is very similar to that of a system of differential equations. In particular, it is still limited to 'cosine' oscillations, though based on the power function instead of the exponential function.

14.9. Exercises

1. By writing the auxiliary equation, show that the solution of $D^2y+aDy=0$ is $y=A+Be^{-ax}$, where A and B are arbitrary. More generally, show that $D^2y+aDy=\phi(x)$ has the same solution as the first-order equation

$$Dy+ay=\psi(x),$$

where $\psi(x)=\int\phi(x)\,dx+A$.

2. If α is a constant, show that a particular integral of $D^2y+aDy+by=\alpha$ is $\bar{y}(x)=\alpha/b$. Generalise for any differential equation with constant coefficients.

3. Write the auxiliary equation of $D^3y+aD^2y+bDy+cy=0$ and exclude the possibility of multiple roots. Show that the differential equation may have a solution which is the sum of three steady exponential terms of form $e^{\lambda x}$. Otherwise, show that it must have one such term and a sinusoidal term. Illustrate by showing that $D^3y-D^2y+2y=0$ has solution

$$y=Ae^{-x}+Be^{x}\cos(x-\epsilon),$$

A, B and ϵ being arbitrary.

*4. *Legendre polynomials.* Consider the differential equation, linear with non-constant coefficients:

$$D^2y-\frac{2x}{1-x^2}Dy+\frac{n(n+1)}{1-x^2}y=0\quad (n \text{ a positive integer}).$$

Show that $y=x$ is a solution when $n=1$, $y=\frac{1}{2}(3x^2-1)$ when $n=2$ and

$$y=\tfrac{1}{2}x(5x^2-3)$$

when $n=3$. Generally, show that $y=P_n(x)$ is a solution, where

$$P_n(x)=\frac{1}{2^n n!}D^n(x^2-1)^n,$$

a polynomial of degree n. These are Legendre polynomials, named after Legendre (1752–1833).

*5. *Bessel functions.* A new function, of a type appearing frequently in mathematical physics and named after Bessel (1784–1846), is defined by the linear differential equation with non-constant coefficients:

$$D^2y+\frac{1}{x}Dy+\left(1-\frac{1}{x^2}\right)y=0.$$

Show that

$$y=\tfrac{1}{2}x-\frac{1}{2!}\left(\tfrac{1}{2}x\right)^3+\frac{1}{2!3!}\left(\tfrac{1}{2}x\right)^5-\ldots+\left(-1\right)^n\frac{1}{n!(n+1)!}\left(\tfrac{1}{2}x\right)^{2n+1}+\ldots$$

is absolutely convergent and satisfies the differential equation ($x\neq0$). (Assume that the series can be differentiated term by term.) This is the expansion of the Bessel function, a case of the hypergeometric series (12.9 Ex. 31).

6. Write the successive derivatives $De^{\lambda x}$, $D^2e^{\lambda x}$, ... and check that

$$F(D)e^{\lambda x}=F(\lambda)e^{\lambda x}$$

where $F(D)$ is a polynomial. Similarly check that

$$F(D)(ye^{\lambda x})=e^{\lambda x}F(D+\lambda)y,$$

where y is a given function of x.

7. If y is a polynomial in x, illustrate how $F(D)y$ can be obtained by showing that

$$(1-D^2)(ax^2+bx+c)=ax^2+bx+(c-2a)$$

and

$$\frac{1}{1-D^2}(ax^2+bx+c)=ax^2+bx+(c+2a).$$

(In the second case, apply $1-D^2$ to the right-hand side.) Also check that

$$\frac{1}{1-D^2}=(1-D^2)^{-1}=1+D^2+D^4+\dots$$

can be applied to ax^2+bx+c with the same result. Solve $D^2y-y+x=0$ by the operator method and check the solution of example (iii) of 14.3.

8. $F(D)y=0$ is a linear differential equation with constant coefficients and $F(D)=0$ has a triple root $D=\lambda$. Show that $y=(A_1+A_2x+A_3x^2)e^{\lambda x}$ is a solution. Generalise to the case of any multiple root.

9. *Resonance.* Show that $D^2y-a^2y=e^x$ has solution $y=Ae^{ax}+Be^{-ax}+\dfrac{e^x}{1-a^2}$, and that $D^2y+a^2y=\cos x$ has solution $y=A\cos(ax-\epsilon)-\dfrac{\cos x}{1-a^2}$. Why does this fail when $a=\pm 1$? This is the phenomenon of resonance.

10. *Simultaneous differential equations.* In general, a pair of linear first-order equations is: $\alpha_{11}Dy+\alpha_{12}Dz+\beta_{11}y+\beta_{12}z=0$ and $\alpha_{21}Dy+\alpha_{22}Dz+\beta_{21}y+\beta_{22}z=0$ If the determinant of the α's is not zero, show that this reduces to form (6) of 14.3, and express the a's of this form in terms of the α's and β's.

11. Show that (6) of 14.3 have an oscillatory solution if $A>\frac{1}{4}(a_{11}+a_{22})^2$, that y and z, while differing in amplitude and phase, have the same period $2\pi/\omega$, where $\omega^2=A-\frac{1}{4}(a_{11}+a_{22})^2$, and that both are similarly damped if $a_{11}+a_{22}<0$.

12. Solve (6) of 14.3 by an alternative method: write $z=ky$ and show that $Dy=\lambda y$, with solution $y=e^{\lambda x}$, $z=ke^{\lambda x}$, where λ and k are given by:

$$a_{11}+ka_{12}=a_{22}+\frac{1}{k}a_{21}=\lambda.$$

Show that there are two values of each of λ and k and complete the solution.

13. *Simultaneous difference equations.* In (8) of 14.6, eliminate z_n and show that:

$$y_{n+2}-(a_{11}+a_{22})y_{n+1}+Ay_n=0$$

where

$$A=\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}.$$

Show that z_n satisfies the same second-order equation. Examine the nature of the solution.

14. Show that $y_{n+2}+ay_{n+1}+by_n=\phi(n)$ in terms of y_n and two subsequent values is equivalent to $y_n+ay_{n-1}+by_{n-2}=\psi(n)$ in terms of y_n and two preceding values, provided that $\psi(n)=\phi(n-2)$.

15. Show that $y_{n+2}+y_{n+1}+\frac{1}{2}y_n=0$ has a solution similar to that of example (iii) of 14.6 but with the shorter period $\frac{8}{3}$. Show that $y_{n+2}-y_{n+1}+y_n=0$ has a solution which is a regular oscillation of period 6.

16. Consider $y_{n+2}-a^2y_n=0$ given y_0 at $n=0$, y_1 at $n=1$. Show that this can be regarded as two independent first-order equations giving $y_n=y_0a^n$ ($n=0$ 2, 4, ...) and $y_n=y_1a^{n-1}$ ($n=1, 3, 5, \dots$). Solve the second-order equation and reconsider the results.

*17. *The shift operator E.* Express a linear difference equation with constant coefficients as $F(E)y_n = \phi(n)$, where $F(E)$ is a polynomial in the shift operator E (defined $Ey_n = y_{n+1}$). Hence get the complementary function from the auxiliary equation $F(E)=0$ and express the particular integral as

$$y_n = \frac{1}{F(E)}\phi(n).$$

*18. Write $E\lambda^n$, $E^2\lambda^n$, $E^3\lambda^n$, ... in terms of λ^n and show that $F(E)\lambda^n = F(\lambda)\lambda^n$. Hence show that the particular integral of $y_{n+2} - y_{n+1} + \frac{1}{2}y_n = (\frac{1}{2})^n$ is $\bar{y}_n = 4(\frac{1}{2})^n$. See example (iii) of 14.6.

19. From the definition, show that the Laplace Transform of e^{ax} is

$$\frac{1}{p-a} \quad (p > a).$$

20. Integrate by parts to show:

$$\int e^{-px}\sin\alpha x\,dx = -\frac{1}{\alpha}e^{-px}\cos\alpha x - \frac{p}{\alpha}\int e^{-px}\cos\alpha x\,dx$$

and
$$\int e^{-px}\cos\alpha x\,dx = \frac{1}{\alpha}e^{-px}\sin\alpha x + \frac{p}{\alpha}\int e^{-px}\sin\alpha x\,dx.$$

Hence obtain the Laplace Transforms of $\sin\alpha x$ and $\cos\alpha x$.

21. Show that $e^{ax}y(x)$ has Laplace Transform $\int_0^\infty e^{-qx}y(x)\,dx = \bar{y}(q)$ where $q = p - a$.

*22. *Laplace Transform for complex p.* Show that the Laplace Transform $\bar{y}(p)$ of $y(x)$ can be defined for complex p, having real part $\int_0^\infty e^{-\alpha x}\cos\omega x\,y(x)\,dx$ and imaginary part $-\int_0^\infty e^{-\alpha x}\sin\omega x\,y(x)\,dx$, where $p = \alpha + i\omega$.

*23. Apply the result of Ex. 21 to show that, if $y_1(x) = e^{ax}\cos bx$ and $y_2(x) = e^{ax}\sin bx$, then $\bar{y}_1(p) = \frac{p-a}{(p-a)^2+b^2}$ and $\bar{y}_2(p) = \frac{b}{(p-a)^2+b^2}$. Write $\alpha = a + ib$ and show that the Laplace Transform of the complex variable $e^{\alpha x} = e^{ax}(\cos bx + i\sin bx)$ is:

$$\frac{(p-a)+ib}{(p-a)^2+b^2} = \frac{1}{p-\alpha}$$

i.e. that the standard form for $e^{\alpha x}$ holds whether α is real or complex.

*24. *Fourier Transform.* Write $Y(p) = \int_0^\infty e^{ipx}y(x)\,dx$, the Fourier Transform of $y(x)$, named after Fourier (1758–1830). Show that $Y(p)$ is a particular case of a Laplace Transform: $Y(p) = \bar{y}(-ip)$. If $y(x) = \sin\alpha x$ and p is real, show that the Laplace Transform $\bar{y}(p) = \frac{\alpha}{\alpha^2+p^2}$ and Fourier Transform $Y(p) = \frac{\alpha}{\alpha^2-p^2}$ are both real.

25. $Z(t) = Ae^{i\epsilon}e^{pt}$ $(p = \alpha + i\omega)$ has real part $z(t) = Ae^{\alpha t}\cos(\omega t + \epsilon)$ and imaginary part $Ae^{\alpha t}\sin(\omega t + \epsilon)$. Show that $e^{-p\theta}$ (θ a real constant) has real

part $e^{-\alpha\theta} \cos \omega\theta$ and imaginary part $e^{-\alpha\theta} \sin \omega\theta$. Deduce that

$$z(t-\theta) = Ae^{-\alpha\theta}e^{\alpha t} \cos (\omega t + \epsilon - \omega\theta)$$

is the real part of $e^{-p\theta}Z(t)$.

26. With $z(t)$ and $Z(t)$ as in Ex. 25, show that $z'(t) = \rho A e^{\alpha t} \cos (\omega t + \epsilon + \eta)$, where $\rho = \sqrt{\alpha^2 + \omega^2}$ and $\tan \eta = \omega/\alpha$, directly from $z(t)$. Then express $Z(t)$ in its real and imaginary parts and write $p = \alpha + i\omega = \rho e^{i\eta} = \rho(\cos \eta + i \sin \eta)$ to show that $z'(t)$ is the real part of $pZ(t)$. Generalise by showing that the nth derivative $z^{(n)}(t) = \rho^n A e^{\alpha t} \cos (\omega t + \epsilon + n\eta)$ and that it is the real part of $p^n Z(t)$, with p^n written $\rho^n e^{in\eta} = \rho^n (\cos n\eta + i \sin n\eta)$.

CHAPTER 15

SOME FORMAL DEVELOPMENT

'God created the natural numbers;
everything else is man's handiwork.'
Kronecker (1823–91).

15.1. From integers to real numbers. (Reference: Chapter 2.) An axiomatic development of the real number system is given here. It includes a brief account of why each particular formulation is expressed in the way chosen. The exposition is carried as far as statements of the main basic properties which follow from the definitions adopted, but the properties are usually not formally established since the proofs are excessively tedious.

(i) *Natural Numbers.* The development here is based on the system of axioms devised by Peano (1858–1932). A 'natural number' is taken as the *primitive (undefined) concept* and the idea of one natural number as the 'successor' of another is a *primitive (undefined) relation.* The object of the exericse is to get a complete set of natural numbers and to arrange them in sequence $1, 2, 3, \dots n, n+1, \dots$. Each natural number is the 'successor' of the one immediately before it in the sequence. In general, $n+1$ is the successor of n. This is, however, hurrying too quickly. Writing '$n+1$' as the successor of 'n' begs the question of what we mean by 'adding 1'. It is better to leave this question over, to be answered at the proper time, and to adopt a neutral symbol 'n^+' for the successor of n.

The following five *axioms* are laid down:*

(1) *Successor:* each natural number n has a unique successor n^+.

(2) *Existence of Natural Numbers:* 1 is a natural number.

* Peano's axioms in themselves characterise any *progression* starting with 1, e.g.

$1, 3, 5, 7, \dots$ where n^+ is 'next odd integer after n'

$1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \dots$ where n^+ is 'half n'.

The particular sequence of natural numbers $(1, 2, 3, 4, \dots)$ follows by the specification of n^+ as $n+1$ and by writing $1+1=2, 2+1=3, \dots$.

(3) *First Natural Number:* no n exists so that $n^+ = 1$.

(4) *Uniqueness:* if $n^+ = m^+$, then $n = m$.

(5) *Completeness:* if a set of natural numbers has the properties: (i) it contains 1, and (ii) it contains n^+ when it contains n, then the set comprises all natural numbers.

Axioms (1), (2) and (3) between them ensure that there is a sequence of natural numbers starting with 1. Though the successor of a number in the sequence is unique by axiom (1), it is possible that two different natural numbers exist with the same successor. This is ruled out by axiom (4). Further, it is still possible that there exist natural numbers outside the sequence. This is ruled out by axiom (5). Hence, by the axioms, the whole set of natural numbers can be written uniquely in sequence. Successive members of the sequence can be given appropriate labels: 1, 2, 3, ... on the established (decimal) notation. Here 2 is a symbol introduced to stand for 1^+, 3 is a symbol for 2^+, and so on.

Axiom (5) which guarantees the completeness of the set (and sequence) of natural numbers is of particular importance. It expresses what is known as the principle of *mathematical induction* and it provides one of the most powerful methods of proof known to mathematicians.

Given the axioms, the essential *definitions* of addition, multiplication and order are:

Sums: to each pair of natural numbers m and n associate a unique natural number, written $m + n$, such that (for any m and n):

$$\text{(i) } n + 1 = n^+ \quad \text{(successor of } n\text{)}$$

and $$\text{(ii) } m + (n + 1) = (m + n)^+ \quad \text{(successor of } m + n\text{)}.$$

Here (i) identifies $n + 1$ as following immediately after n in the sequence of natural numbers, i.e. $2 = 1 + 1$ since 2 follows 1; $3 = 2 + 1$ since 3 follows 2; and so on. Then (ii) serves to reverse the process, enabling several steps to be taken in any order. For example, since 2 is $1 + 1$, 3 is the successor of $1 + 1$ which by (ii) is $1 + (1 + 1) = 1 + 2$. Hence, $2 + 1 = 1 + 2$ and equally $(1 + 1) + 1 = 1 + (1 + 1)$, all being 3. Generally, $m + n = n + m$ and $(m + n) + p = m + (n + p)$ for any m, n and p. These are the commutative and associative rules.

Products: to each pair of natural numbers m and n associate a unique natural number, written $m \times n$, such that (for any m and n):

$$\text{(i) } n \times 1 = n \quad \text{and} \quad \text{(ii) } m \times (n+1) = m \times n + m.$$

Products are identified as repeated sums:

$$n \times 1 = n; \quad n \times 2 = n \times (1+1) = n \times 1 + n = n + n;$$
$$n \times 3 = n \times (2+1) = n \times 2 + n = n + n + n;$$

and so on. It then follows, by bringing in the rules already established for sums, that similar commutative and associative rules hold for products: $mn = nm$ and $(mn)p = m(np)$. (The sign × can always be dropped when there is no ambiguity.) There is, further, a relation between sums and products. For example: $m(n+1) = mn + m$ by (ii); another application of (ii) gives

$$m(n+2) = m(n+1+1) = m(n+1) + m = mn + m + m = mn + m \times 2;$$

and so on. In general, $m(n+p) = mn + mp$ for any m, n and p, the distributive rule.

Order: if m and n are such that $m = n + p$ for some p, then n is said to be less than m, written $n < m$. In particular, since $n^+ = n + 1$, this definition of order implies that $n < n^+$. A natural number n is less than its successor $n + 1$. The basic order is 1, 2, 3, 4, ...

It follows almost immediately from these definitions that the set J^+ of the natural numbers satisfies all the operational rules and order properties of 2.2, *except* that zero, negatives and reciprocals are lacking and that the order is not dense and not extended in both directions.

(ii) *Integers*. From the set J^+ of natural numbers (positive integers) a wider set J of all integers (positive, negative, zero) is defined.† The idea is that, given two natural numbers n and m, they define a positive integer by the 'difference' $m - n$ if $n < m$, they define zero by the 'difference' $m - n = 0$ if $n = m$, and they define a negative integer by the 'difference' $m - n = -(n - m)$ if $n > m$. However, the question-begging word 'difference' and notation $m - n$ must be avoided in the definition. This is done by writing the pair (m, n) of natural numbers

† Historically, this was not the first step. It was easier to go from the natural numbers to ratios p/q of natural numbers, i.e. to positive rationals. The concept of negative numbers came later and it was not readily accepted. One difficulty was in the *rule of signs*, still puzzling to children:

$$(+1) \times (+1) = +1; \ (-1) \times (-1) = +1; \ (+1) \times (-1) = -1; \ (-1) \times (+1) = -1.$$

The definition given here exhibits the rule of signs in the form:

$$(1, 0) \times (1, 0) = (1, 0); \ (0, 1) \times (0, 1) = (1, 0); \ (1, 0) \times (0, 1) = (0, 1); \ (0, 1) \times (1, 0) = (0, 1).$$

instead of $m-n$. Further, a difficulty must be overcome: there are many pairs with the same 'difference'. For example, to get the negative integer -2, we have $1-3=2-4=3-5=\ldots=-2$. The safe way of writing -2 is as the number pair (m, n) with $m+2=n$. So:

DEFINITION: *An* **integer** *is the pair of natural numbers* (m, n) *such that* $+$, $\times$ *and* $<$ *are defined for* (m, n) *and* (p, q):

$$(m, n)+(p, q)=(m+p, n+q)$$
$$(m, n)\times(p, q)=(mp+nq, mq+np)$$
$$(m, n)<(p, q) \text{ if } m+q<n+p.$$

Then (m, n) *is the* **positive integer** $m-n$ *if* $n<m$, *the integer* **zero** *if* $n=m$, *and the* **negative integer** $-(n-m)$ *if* $n>m$.

The definitions of $+$, $\times$ and $<$ appear artificial. They are designed to carry these operations over from natural numbers to integers. For example:

$$(m, n)=-2 \text{ if } m+2=n; \quad (p, q)=3 \text{ if } p=q+3.$$

So:
$$\begin{aligned}(-2)\times 3 &= (m, n)\times(p, q)=(mp+nq, mq+np)\\ &= \{m(q+3)+(m+2)q, mq+(m+2)(q+3)\}\\ &= (2mq+3m+2q, 2mq+3m+2q+6)\\ &= (r, s) \quad \text{where } r+6=s \quad (r-s=-6)\\ &= -6.\end{aligned}$$

The set J^+ of natural numbers is extended to the wider set J of integers; amongst the elements of J are the positive integers, the equivalents of the natural numbers of J^+. It follows quite easily from the definition that J satisfies all the operational rules and order properties of 2.2, *except* that reciprocals are still lacking and that the order is not dense. The set J is an integral domain.

(iii) *Rational Numbers.* One further step is needed to eliminate the lack of reciprocals in J. It is a rather general process, known as the definition of a *quotient field.* In this case, if p and q are integers of J, the quotient p/q is formed to provide the system of rationals, a field of quotients. At the same time, the definition makes good the deficiency in the order; the rationals have all the order properties, including that of density.

In the definition of rational numbers, the idea is to write the 'ratio' or 'quotient' of one integer by another: $p \div q$ or p/q. To avoid question-begging, these terms and notations cannot be used in the definition. There is again the difficulty that many different pairs of integers have the same quotient, e.g. the rational number $\frac{2}{3}$ is got equally from $\frac{4}{6} = \frac{6}{9} = \dots$. The safe definition, therefore, is to write a rational number as a pair of integers (m, n) with the convention added that $(m, n) = (p, q)$ if the integers mq and np are equal.†

DEFINITION: *A* **rational number** *is the pair of integers* (m, n), *where* $(m, n) = (p, q)$ *if* $mq = np$, *such that* $+$, $\times$ *and* $<$ *are defined:*

$$(m, n) + (p, q) = (mq + np, nq)$$

$$(m, n) \times (p, q) = (mp, nq)$$

$$(m, n) < (p, q) \text{ if } mq < np \quad \text{for } n \text{ and } q \text{ positive.}$$

Then (m, n) *is written* $\frac{m}{n}$ $(n \neq 0)$, $\frac{m}{n} = \frac{p}{q}$ *if* $mq = np$ *and* $\frac{m}{n} < \frac{p}{q}$ *if* $mq < np$ (*n and q positive*).

The reason for adopting the particular definitions for $+$ and $\times$ is clear in translation into the $\frac{m}{n}$ notation: $\frac{m}{n} + \frac{p}{q} = \frac{mq + np}{nq}$ and $\frac{m}{n} \times \frac{p}{q} = \frac{mp}{nq}$. To be definite on 'less than' ($<$), it can always be arranged that n and q are positive integers (in the denominators): $\frac{m}{n} < \frac{p}{q}$ if $mq < np$.

It follows easily from the definition that rational numbers satisfy all the operational rules and order properties of 2.2. The set R of rationals is an ordered field, containing within it the particular rationals $\frac{m}{1}\left(\frac{m}{n} \text{ with } n = 1\right)$ which are equivalent to the integers m.

(iv) *Real Numbers.* The development of the number system into a complete ordered field is rounded off by the definition of the set R^* of real numbers. The need for completion appears on splitting the set R of rationals into subsets L and G so that each rational is in L or in G, each of L and G contains at least one rational, and each rational of L is less than each rational of G. This is called a *Dedekind Cut,*

† The condition is written, as it must be, in terms of integers only, i.e. $mq = np$. It will appear as $m/n = p/q$ in the notation to be adopted.

after Dedekind (1821–1916). There are three possibilities:† (i) L has a greatest member a and G a least member b; (ii) L has a greatest member a, or G a least member b, but not both; (iii) L has no greatest member and G no least. Case (i) can be eliminated. If it holds, then $a<b$ by definition of the cut and so $\frac{1}{2}(a+b)$ is a rational which is in neither L nor G, a contradiction. The other two cases remain; examples are: (ii) L given as all rationals $x \leqslant 2$ and G as all rationals $x>2$; (iii) L given as all negative rationals and positive rationals x such that $x^2 \leqslant 2$ and G as all positive rationals x such that $x^2>2$.

Hence, if we say that a Dedekind cut *defines* a number (as the dividing point between L and G), the number may correspond to a rational, but it may not. This suggests the definition of the wider set R^* of real numbers:

DEFINITION: *A Dedekind Cut of the set R of rationals into subsets L and G so that*

(1) *each rational is in L or in G*

(2) *each of L and G contains at least one rational, and*

(3) *each rational of L is less than each rational of G*

defines a **real number** *α as the point of division between L and G.*

Hence, R^* includes some α corresponding to rational numbers; other α of R^* do not so correspond and they are *irrational* real numbers. As an example of an irrational real number (apart from $\surd 2$ already quoted), consider the equation $x^3-3x-8=0$. If $x=a$ is a rational root, then $(x-a)$ is a factor of x^3-3x-8 and a is an integer dividing 8; the only possibilities are $a=1$, 2, 4 or 8 and none satisfies the equation. There is no rational root. A Dedekind Cut is given by $x^3-3x-8 \leqslant 0$ (x in L) and $x^3-3x-8>0$ (x in G) as can be seen by plotting a graph of $y=x^3-3x-8$ for rational x. The real number α so defined is an irrational root of $x^3-3x-8=0$.

From the definition it follows that real numbers do as well, algebraically, as the rationals; they satisfy all the rules and properties of 2.2, making R^* an ordered field. The important matter is the redefinition of $+$, $\times$ and $<$:

† In case (i) there is a *jump* from L to G; in case (iii) a *gap* between L and G. A cut in the set of integers gives rise to case (i), i.e. integers have jumps. A cut in the set of rationals excludes (i) but (iii) is possible, i.e. rationals have gaps but no jumps. For real numbers, only case (ii) arises and there are neither jumps nor gaps. This is what we mean when we say that real numbers form a 'continuum'.

Let α' be a real number given by a Dedekind Cut with L' containing rationals a' and G' rationals b', where $a' < b'$. Let α'' be another real number, the subsets of the Dedekind Cut being L'' (rationals a'') and G'' (rationals b'') where $a'' < b''$.

Addition of α' and α'': write $a = a' + a''$, $b = b' + b''$ as sums of rationals. Then all a make up a subset L, and all b a subset G. Since $a < b$, L and G give a Dedekind Cut and the real number so specified is defined as $\alpha = \alpha' + \alpha''$.

Multiplication of α' and α'': if $\alpha' \geqslant 0$ and $\alpha'' \geqslant 0$ (i.e. G' and G'' consist of positive rationals only), then the Dedekind Cuts can be limited to cuts of the non-negative rationals and there is no difficulty on signs in multiplying rationals. Write $a = a' \times a''$ making up a subset L, and write $b = b' \times b''$ making up a subset G, of the non-negative rationals. Since $a < b$, L and G give a Dedekind Cut of the non-negative rationals and the real number which corresponds is defined as $\alpha = \alpha' \times \alpha''$. If $\alpha' < 0$ and/or $\alpha'' < 0$, then $\alpha = \alpha' \times \alpha''$ is defined in terms of the corresponding non-negative numbers, e.g. $\alpha'\alpha'' = -\alpha'(-\alpha'')$ if $\alpha' > 0$, $\alpha'' < 0$.

Order of α' and α'': if α' and α'' are different (i.e. L' and L'' are different subsets, as are G' and G''), then $a' < b'$ and $a'' < b''$ imply only two possibilities. Either all a' of L' are in L'' (and so all b'' of G'' are in G') in which case we define $\alpha' < \alpha''$; or all a'' of L'' are in L' (and so all b' of G' are in G'') in which case we define $\alpha'' < \alpha'$.

Once these re-definitions are made, it is a straight-forward but exceedingly tedious task to check that all the rules and properties of 2.2 hold for real numbers. This task is not undertaken here. The important result is:

THEOREM: *If a Dedekind Cut* (L *and* G) *is made of the set* R^* *of real numbers, then it gives a unique real number* α *such that all real numbers* $a < \alpha$ *are in* L *and all real numbers* $b > \alpha$ *are in* G, *and such that* α *belongs either to* L *or to* G.

Proof: Let L' and G' be respectively the set of *rationals* in L and G, which are subsets of real numbers. There are two possibilities:

(i) L' has a greatest member, the rational a. L may still contain a real number $x > a$. Then the Dedekind Cut of rationals giving x must contain in its lower subset some rationals between a and x; a contradiction. Hence L has no real member greater than a, i.e. L has a as its greatest member. A similar situation arises if G' has a least member, so that G has a least member, a rational b.

(ii) L' has no greatest and G' no least member. Then the Dedekind Cut of rationals (comprising L' and G') gives a real number α which is irrational, and which belongs either to L or to G. If α belongs to L, it must be L's greatest member; otherwise, there is a real $x > \alpha$ in L and so rationals between α and x belonging to L', contradicting the condition that α is the real number given by L' and G'. Similarly, if α belongs to G, it must be G's least member. It follows, in both cases (i) and (ii), that there is a real number (rational or irrational) which is either the greatest of L or the least of G. It must be unique; otherwise, if α' and α'' are one the greatest of L and the other the least of G, then $\frac{1}{2}(\alpha' + \alpha'')$ is neither in L nor in G, a contradiction.

Q.E.D

The implication of the theorem is most important. A Dedekind Cut of *rational* numbers gives either a rational or something new (an irrational). The same process applied to *real* numbers produces nothing new; a Dedekind Cut of real numbers always gives a real number, belonging to one or other of the two sets L and G of the Cut. The set R^* of real numbers is complete, cut it which way we will. It follows that, if a set S of real numbers with a lower bound is given, then a Dedekind Cut can be specified: x belongs to L if it is a lower bound of S, otherwise x belongs to G. The real number so defined by the Cut is the GLB of S. As such, it may or may not belong to S. A similar result holds for a LUB.

15.2. Polynomials: the fundamental theorem of algebra. (Reference: Chapter 3.) The essential idea of a polynomial is that it is a sequence of coefficients, drawn from a given field F. In writing polynomials with rational coefficients, as in 3.3, F is the field R of rationals. F could equally well be the field of real (or complex) numbers giving polynomials with real (or complex) coefficients.†

Definition: *A* **polynomial** *over the field F is any sequence $(f_0, f_1, f_2, f_3, \ldots)$ of elements of F with only a finite number of non-zero terms, and subject to the rules of addition and multiplication:*

† It is also possible to take F as an integral domain, e.g. the set of integers, giving polynomials with integral coefficients. The particular expression (7) below does not then apply.

$$\left.\begin{aligned}&(f_0, f_1, f_2, f_3, \ldots)+(g_0, g_1, g_2, g_3, \ldots)\\&\quad=(f_0+g_0, f_1+g_1, f_2+g_2, f_3+g_3, \ldots)\end{aligned}\right\} \quad \ldots\ldots\ldots\ldots(1)$$

$$\left.\begin{aligned}&(f_0, f_1, f_2, f_3, \ldots)\times(g_0, g_1, g_2, g_3, \ldots)\\&\quad=(h_0, h_1, h_2, h_3, \ldots)\\ \textit{where}\quad & h_m=f_0g_m+f_1g_{m-1}+\ldots+f_mg_0\end{aligned}\right\} \quad \ldots\ldots\ldots\ldots(2)$$

As a convention, identify the polynomial $(f_0, 0, 0, 0, \ldots)$ with the element f_0 of F, e.g. with the rational f_0 when F is the field R. Then:

$$\begin{aligned}&(f_0, f_1, f_2, f_3, \ldots)\\&\quad=(f_0, 0, 0, 0, \ldots)+(0, f_1, 0, 0, \ldots)+(0, 0, f_2, 0, \ldots)+\ldots \quad \text{by (1)}\\&\quad=(f_0, 0, 0, 0, \ldots)+(f_1, 0, 0, 0, \ldots)\times(0, 1, 0, 0, \ldots)\\&\qquad\quad+(f_2, 0, 0, 0, \ldots)\times(0, 0, 1, 0, \ldots)+\ldots \quad \text{by (2)}\end{aligned}$$

i.e. $(f_0, f_1, f_2, f_3, \ldots)$

$$=f_0+f_1(0, 1, 0, 0, \ldots)+f_2(0, 0, 1, 0, \ldots)+\ldots \quad \ldots\ldots\ldots\ldots\ldots\ldots(3)$$

Write x for the particular polynomial $(0, 1, 0, 0, \ldots)$. Apart from the fact that it is one of the polynomials, we still leave x undefined. Then:

$$x^2=(0, 1, 0, 0, \ldots)\times(0, 1, 0, 0, \ldots)=(0, 0, 1, 0, \ldots) \quad \text{by (2)}$$

$$x^3=(0, 0, 1, 0, \ldots)\times(0, 1, 0, 0, \ldots)=(0, 0, 0, 1, 0, \ldots)$$

and so on. Hence x^m, the product of x by itself (m times) is itself a polynomial, that with 1 in the mth place and zero elements elsewhere. Substituting in (3), we have the familiar notation for a polynomial:

$$(f_0, f_1, f_2, f_3, \ldots)=f_0+f_1x+f_2x^2+f_3x^3+\ldots \quad \ldots\ldots\ldots\ldots(4)$$

Since there is only a finite number of non-zero elements among the f's, there is a last non-zero element. Let it be f_n and stop the sequence here. This means that there can be zero values for any or all of $f_0, f_1, f_2, \ldots f_{n-1}$, that $f_n \neq 0$, and that any subsequent elements can be ignored as all zero. Hence the sequence of *coefficients* $f_0, f_1, f_2, \ldots f_n$ is obtained, for $n \geqslant 0$ and $f_n \neq 0$. The polynomial is then of *nth degree* and it can be written from (4):

$$f(x)=f_0+f_1x+f_2x^2+\ldots+f_nx^n \quad (n\geqslant 0, f_n\neq 0) \quad \ldots\ldots\ldots(5)$$

The set of all polynomials (5), as n and the f's vary, is denoted $F[x]$.

A further development is possible if we agree to impose another rule on the set of polynomials. This defines *scalar multiplication*, i.e. multiplication of $f(x)$ by any element λ of the field F:

$$\lambda(f_0, f_1, f_2, f_3, \ldots)=(\lambda f_0, \lambda f_1, \lambda f_2, \lambda f_3, \ldots) \quad \ldots\ldots\ldots\ldots\ldots\ldots(6)$$

The rule (6), which is additional to (1) and (2), applies to (5):

$$\lambda(f_0+f_1x+f_2x^2+\ldots+f_nx^n)=(\lambda f_0)+(\lambda f_1)x+(\lambda f_2)x^2+\ldots+(\lambda f_n)x^n.$$

Put $\lambda=\frac{1}{f_n}$ as a particular case, λ being an element of F:

$$\frac{1}{f_n}f(x)=\frac{f_0}{f_n}+\frac{f_1}{f_n}x+\frac{f_2}{f_n}x^2+\ldots+\frac{f_{n-1}}{f_n}x^{n-1}+x^n$$

i.e. $$g(x)=x^n+a_{n-1}x^{n-1}+\ldots+a_2x^2+a_1x+a_0 \quad\ldots\ldots\ldots\ldots(7)$$

where $a_r=\frac{f_r}{f_n}$, elements of F for $r=0, 1, 2, \ldots n-1$, and where $g(x)=\frac{1}{f_n}f(x)$. For most purposes $g(x)$ is as good as $f(x)$; both are polynomials and they differ only by a constant factor. Hence, (7) can usually be taken as the general expression of a polynomial of degree n.

The rules (1) and (2), for sums and products, when applied to the general polynomial of form (5), become the familiar processes of elementary algebra:

If $f(x)=f_0+f_1x+f_2x^2+\ldots+f_nx^n$ and $g(x)=g_0+g_1x+g_2x^2+\ldots+g_mx^m$, then

$$\left.\begin{aligned}f(x)+g(x)&=(f_0+g_0)+(f_1+g_1)x+(f_2+g_2)x^2+\ldots\\ f(x)\times g(x)&=f_0g_0+(f_0g_1+f_1g_0)x+(f_0g_2+f_1g_1+f_2g_0)x^2+\ldots\end{aligned}\right\}\ldots\ldots(8)$$

The sum polynomial terminates with f_nx^n if $n>m$, with $(f_n+g_n)x^n$ if $n=m$ or with g_mx^m if $n<m$. The product polynomial terminates with $f_ng_mx^{n+m}$. From (8), it follows that polynomials $f(x)$ of the set $F[x]$ satisfy all the operational rules, except that reciprocals are not defined. $F[x]$ is an integral domain.

The lack of reciprocals can be rectified, as in 15.1 (iii), by constructing the quotient field of $F[x]$, i.e. the set of ratios of polynomials.

DEFINITION: *A* **rational fraction** $\frac{f(x)}{g(x)}$ *is a pair of polynomials* $\{f(x), g(x)\}$, $g(x)\neq 0$, *such that* + *and* × *are defined:*

$$\{f(x), g(x)\}+\{\phi(x), \psi(x)\}=\{f(x)\psi(x)+g(x)\phi(x), g(x)\psi(x)\}$$

$$\{f(x), g(x)\}\times\{\phi(x), \psi(x)\}=\{f(x)\phi(x), g(x)\psi(x)\}.$$

The rational fraction $\frac{f(x)}{f(x)}$ is identified as 1, i.e. $f(x)$ cancels out. The rational fraction $\frac{f(x)}{1}$ is identified as $f(x)$, i.e. polynomials are included in the set of rational fractions. To distinguish it from the set $F[x]$ of polynomials, the set of rational fractions is denoted $F(x)$.

The reciprocal of $f(x)$ can now be written as $\frac{1}{f(x)}$. For:

$$f(x)\times\frac{1}{f(x)}=\{f(x),\,1\}\times\{1,f(x)\}=\{f(x),f(x)\}=\frac{f(x)}{f(x)}=1.$$

The set $F(x)$ of all rational fractions satisfies the whole system of operational rules of 2.2. While $F[x]$ is only an integral domain, $F(x)$ is a field. There is no question of ordering rational fractions. $F(x)$ is not an ordered field.

The method of extending a given field F into a wider field $F(x)$ of rational fractions $f(x)/g(x)$ is known as *adjunction* of x to the field F. Here x is some element outside the given field F. First form sums of products:

$$f(x)=f_0+f_1x+f_2x^2+\ldots+f_nx^n \quad (n\geqslant 0),$$

by taking x both with itself and with elements of F, subject to a redefinition of addition $(+)$ and multiplication $(\times)$ as given by (8). The result is an integral domain of elements $f(x)$. Next, write $f(x)/g(x)$ where $f(x)$ and $g(x)$ are both of the form indicated, i.e. where

$$\frac{f(x)}{g(x)}=\frac{f_0+f_1x+f_2x^2+\ldots+f_nx^n}{g_0+g_1x+g_2x^2+\ldots+g_mx_m} \quad (n\geqslant 0,\ m\geqslant 0,\ \ g_m\neq 0) \quad \ldots\ldots\ldots(9)$$

The set of elements (9) is a field, the adjunction of x to F. The notation $F(x)$ for this field indicates the adjunction. The original field F has been expanded to a wider field $F(x)$ by swallowing up x together with all the necessary sums, products and quotients.

Adjunction is a very powerful tool for operating on fields. A simple example: take the field R of rationals and form the field $R(\surd 2)$ by adjunction of the outside element $\surd 2$. Put $x=\surd 2$ in (9) and note that $x^2=2$, $x^3=2\surd 2$, $x^4=4$, ... The general element of $R(\surd 2)$ is then of the form:

$$\frac{a+b\surd 2}{c+d\surd 2}=x+y\surd 2$$

by multiplying numerator and denominator by $c-d\surd 2$. This is done in 2.3.

A more important example is to write R^* (i) by adjunction of the element i to the field R^* of real numbers, where $i^2=-1$. Put $x=i$ in

(9), noting that $x^2 = -1$, $x^3 = -i$, $x^4 = 1$, ... The general element of $R^*(i)$ is:

$$\frac{a+ib}{c+id} = x+iy$$

by multiplying numerator and denominator by $c-id$. Hence, as indicated in 2.5, the field C of complex numbers is obtained by adjunction of i to R^*.

It remains to establish the *Fundamental Theorem of Algebra* that every polynomial $f(x)$ of positive degree, over the field F of rational, real or complex numbers, is such that $f(\alpha)=0$ for some complex number α. The consequence of this theorem (as shown in 3.7) is that a polynomial of degree $n>0$ has precisely n linear factors, and the polynomial equation has precisely n roots. The theorem is stated in 3.7 for a polynomial with rational coefficients. It is equally true for polynomials with real or complex coefficients.

THEOREM: *If $f(x)$ is any polynomial with rational, real or complex coefficients, and of positive degree, then $f(\alpha)=0$ for some complex α.*
All known proofs of this theorem are framed in terms which involve topological concepts. It would appear to be incapable of proof in purely algebraic terms. The following is no more than a sketch of a proof, the details and refinements of the topological construction being omitted.

Proof: Write $f(x) = x^n + a_{n-1}x^{n-1} + \dots + a_1x + a_0$, where n is the degree of the polynomial. Let $\alpha = re^{i\theta}$ be any complex number, where r is the absolute value and θ the argument (12.7 above), represented by a point P on an Argand Diagram. Fig. 15.2 illustrates. As α varies, so do r and θ and P moves around the plane. We can get α to cover all complex values, and P to move over the whole plane, by first letting P describe an anti-clockwise circle (given r, θ increasing from 0 to 2π) and then increasing the radius of the circle from O outwards (r increasing from 0).

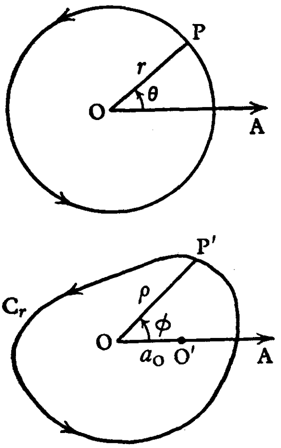


FIG. 15.2

Write $\beta = f(\alpha)$, a function of a complex variable α. Given $\alpha = re^{i\theta}$, then β is a specific complex number $\beta = \rho e^{i\phi}$, where ρ is its absolute value and ϕ its argument, represented by a point P' corresponding to P (Fig. 15.2). Then ρ and ϕ vary continuously with r and θ. As P describes some continuous curve, then P' describes a continuous curve. The relation $\beta=f(\alpha)$ maps P onto P', maps one curve onto the other. Let P describe an anti-clockwise circle of radius r and let C_r be the corresponding curve described by P'. Write m for the net number of times C_r goes round O anti-clockwise, net in the sense that a clockwise turn = a negative turn. So m depends on r, giving a continuous function $m(r)$.

When $r=0$, then $\alpha=0$ and the 'circle' is the point O. But

$$\beta=f(0)=a_0,$$

i.e. the 'curve' C_r $(r=0)$ is the single point O' (corresponding to O). Hence $m(0)=0$. When r is large, P describes a large circle (α large). Then $\beta=f(\alpha)=\alpha^n=r^ne^{ni\theta}$ approximately from the polynomial. As P goes round its circle (θ increasing from 0 to 2π), P' is given by $\beta=r^ne^{ni\theta}$, i.e. $\rho=r^n$ (fixed) and $\phi=n\theta$ (increasing from 0 to $2n\pi$). Here, as P' describes C_r, it goes n times round O. So $m(r)\to n$ as $r\to\infty$. The curve C_r starts as a single point $(r=0)$ and finishes by winding itself n times round O $(r\to\infty)$. At some r (and for some θ), C_r must go through O. When it does, $\beta=0$. The value of α which corresponds is such that $f(\alpha)=0$. Q.E.D.

15.3. Sets, groups, fields and vector spaces. (Reference: Chapters 4, 6 and 8.)

(i) *Set Theory.* The axiomatic development of the theory of sets, as a foundation for all mathematics, can be achieved in various ways. The following is based on Zermelo's system of axioms. The *primitive (undefined) concepts* are those of a 'set' and of the relation 'belongs to'. The notation $a \in A$ indicates that the object a belongs to the set A. As consequent *definitions:*

(*a*) A is a *subset* of B, denoted $A \subseteq B$, if $a \in A$ implies $a \in B$ for all a.

(*b*) A *equals* B, denoted $A=B$, if $A \subseteq B$ and $B \subseteq A$ both hold.

Further, A is a *proper subset* of B if $A \subseteq B$ holds and if $A=B$ does not.

The following five *axioms* are laid down:

(1) *Members:* a set is fully determined by its members so that, if a and b are equal (identical) objects, then $a \in A$ implies $b \in A$.

(2) *Pairing:* if a and b are different objects, then there exists a set $\{a, b\}$ comprising just a and b, and no more.

(3) *Selection:* if the property Π is meaningful for members of a set A, then there exists a subset of A comprising just those members of A for which Π holds, and no more.

(4) *Sum Set:* if the set X comprises the sets $A, B, C, \ldots$, then there exists a sum set $SX = A + B + C + \ldots$ comprising just the members of $A, B, C, \ldots$, and no more.

(5) *Power Set:* if the set X has subsets $A, B, C, \ldots$, then there exists a power set $PX = \{A, B, C, \ldots\}$ comprising all subsets of X.

Axiom (1) ensures that a set A can be denoted by its members $\{a, b, c, \ldots\}$, where the order of writing the members is immaterial and where the number of members need not be finite. Axiom (2), in which the 'objects' a and b can themselves be sets, is designed to build up sets from individual objects or from smaller sets.

A property is 'meaningful' if it is true or false (and not irrelevant) for members of A; for example the property 'a is even' is meaningful for the set of natural numbers but not for a set of persons. Axiom (4) relates to sets $A, B, C, \ldots$ which may be single objects as well as finite or infinite sets. If $A, B, C, \ldots$ are all the sets formed from the totality of elements in a *universal set* U, then $\{A, B, C, \ldots\}$ exists as the power set of U by axiom (5).

A sequence of *further definitions* is obtained from the axioms. The *nul set* ϕ is the set comprising no member and the *unit set* $\{a\}$ of an object a is the set comprising just a and no more. The existence of these sets is guaranteed by Axioms (2) and (3). If a and b are different objects, write $S = \{a, b)$ as the pair set and take property Π as 'not a member of S', giving a subset of S which is the nul set. If Π is taken as 'different from b', it gives a subset of S which is the unit set $\{a\}$.

Given two sets A and B, the *union* $A \cup B$ is the set comprising just those members belonging either to A or to B. The union exists in virtue of Axiom (4). The *intersection* $A \cap B$ is the set comprising just those members belonging both to A and to B. Its existence is guaranteed by Axiom (3) with Π as '$a \in B$' taken over all members a of A. The *Cartesian product* $A \,.\, B$ is the set comprising all pairs (a, b) where $a \in A$ and $b \in B$. It exists because of Axiom (5), with the power set $P(A \cup B)$ formed from the union $A \cup B$. Take Π as '$S \cap A$ and $S \cap B$ are each unit sets' and apply to members S of $P(A \cup B)$. Axiom (3) then gives $A \,.\, B$ as a subset of $P(A \cup B)$. In other words, $A \cup B$ has a variety of subsets S and Π picks out those consisting of two elements, one from A (given by $S \cap A$) and one from B (given by $S \cap B$).

Consider $\{A, B, C, \ldots\}$ as the power set of the universal set U. One member is the nul set ϕ. In addition to the *union* $A \cup B$ and the *intersection* $A \cap B$ for any pair A and B, there can be defined the *complement* A' of any set A, comprising just those members of U not belonging to A. A rather complete set of properties is:

Rule	*Union* ($\cup$)	*Intersection* ($\cap$)
Closure	$A \cup B \in U$	$A \cap B \in U$
Associative	$A \cup (B \cup C) = (A \cup B) \cup C$	$A \cap (B \cap C) = (A \cap B) \cap C$
Commutative	$A \cup B = B \cup A$	$A \cap B = B \cap A$
Idempotent	$A \cup A = A$	$A \cap A = A$
Bounds	$A \cup \phi = A$ $A \cup U = U$	$A \cap U = A$ $A \cap \phi = \phi$
Distributive	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
Complements	$A \cup A' = U$ $(A \cup B)' = A' \cap B'$	$A \cap A' = \phi$ $(A \cap B)' = A' \cup B'$
Involution	$(A')' = A$	

The proofs are straight-forward and they are assisted by drawing Venn Diagrams. The distributive rule

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

is illustrated in Fig. 15.3. The *total* shaded area in the top diagram represents the left-hand set; the *cross* shaded area in the bottom diagram represents the right-hand set. The two areas are the same.

A *finite set* is a set which can be put into one-one correspondence with $\{1, 2, 3, \ldots n\}$ for some natural number n. Any other set is an *infinite set*. There is nothing in the five axioms above to guarantee the existence of any infinite set. This requires an additional axiom:

(6) *Infinite sets:* there is a set, which is infinite, comprising all natural numbers. Other infinite sets can be built up once this particular set is known.

A set is *reflexive* if it can be put into one-one correspondence with a proper subset of itself. To establish that any infinite set is reflexive requires a further additional axiom.

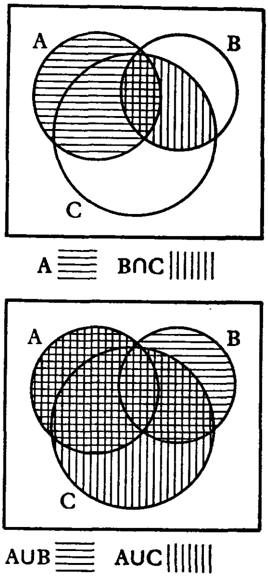


FIG. 15.3

This relates to the Cartesian product, extended from the case of $A \,.\, B$ for two sets to that of $A \,.\, B \,.\, C \,.\, \ldots$ for any set of sets $\{A, B, C, \ldots\}$. For example, with three sets, $A \,.\, B \,.\, C$ is the set of all triples (a, b, c) where $a \in A$, $b \in B$ and $c \in C$. The axiom is:

(7) *Axiom of Choice:* if the sets $A, B, C, \ldots$ are each non-empty and pairwise disjoint (no common element), then the Cartesian product $A \,.\, B \,.\, C \,.\, \ldots$ is non-empty.

It can be shown that the axiom guarantees that, given a set X and hence its power set $PX = \{A, B, C, \ldots\}$ comprising all subsets, a definite element of each non-empty $A, B, C, \ldots$ can be picked out. In its turn, this guarantees that a countably infinite sequence $x_1, x_2, x_3, \ldots$ can be selected from any infinite set X. The successive subsets of PX taken for the selection of definite elements are X itself, X with x_1 removed, and so on. It then follows that all infinite sets are reflexive.†

Algebra is largely concerned with sets which have the structure of a group. Development proceeds in two directions, to specialised sets with the structure of a ring and to other specialised sets with the structure of a vector space. These classifications are not exclusive; for example, a field is a special form of a ring and it may have properties which qualify it as a vector space.

(ii) *Groups.* With an operation $*$, the definition of a group is:

DEFINITION: *A set G is a* **group** *if, for all elements $a, b, c, \ldots$:*

(1) *Closure:* $a * b \in G$

(2) *Associative:* $a * (b * c) = (a * b) * c$

(3) *Identity: there is an $e \in G$ so that $e * a = a$*

(4) *Inverse: there is an $a^{-1} \in G$ so that $a^{-1} * a = e$.*

The following properties are established from the definition:

(a) If a^{-1} is an inverse of a, so that $a^{-1} * a = e$, then $a * a^{-1} = e$.

For:

$$\begin{aligned}(b * a^{-1}) * (a * a^{-1}) &= b * (a^{-1} * a) * a^{-1} = b * e * a^{-1} && \text{by (2) and (4)}\\ &= b * a^{-1} && \text{by (3).}\end{aligned}$$

† The seven axioms, including the Axiom of Choice, can be shown to be consistent one with another. They are not necessarily complete for all types of sets used in set theory. On the axiomatic basis of set theory, see Fraenkel: *Abstract Set Theory* (North-Holland, 1953) and Fraenkel and Bar-Hillel: *Foundations of Set Theory* (North-Holland, 1958).

Write b as an inverse of a^{-1}, so that $b * a^{-1} = e$ by (4).

Then: $e * (a * a^{-1}) = e$

i.e. $a * a^{-1} = e$ by (3). Q.E.D.

(*b*) If e is an identity, so that $e * a = a$, then $a * e = a$.

For: $a * e = a * (a^{-1} * a) = (a * a^{-1}) * a$ by (4) and (2)

$= e * a = a$ by (*a*) Q.E.D.

(*c*) If $a * b = a * c$, then $b = c$.

For: from the given data, $a^{-1} * (a * b) = a^{-1} * (a * c)$.

Now: $a^{-1} * (a * b) = (a^{-1} * a) * b = e * b = b$ by (2), (4) and (3).

Similarly: $a^{-1} * (a * c) = c$. Hence, $b = c$. Q.E.D.

(*d*) The identity e is unique, and so is the inverse a^{-1} of a.

For: if there are two identities e and e', then

$a * e = a * e'$ by (*b*)

i.e. $e = e'$ by (*c*).

The uniqueness of a^{-1} is proved similarly. Q.E.D.

Note: these four properties complete the table of the operational rules for groups, given in 6.2.

(*e*) $(a * b)^{-1} = b^{-1} * a^{-1}$.

For: $(b^{-1} * a^{-1}) * (a * b) = b^{-1} * (a^{-1} * a) * b = b^{-1} * e * b$

$= b^{-1} * b = e$

i.e. $b^{-1} * a^{-1}$ is the inverse of $a * b$. Q.E.D.

(*f*) If $b * x = a$, then $x = b^{-1} * a$; if $y * b = a$, then $y = a * b^{-1}$.

For: $b * (b^{-1} * a) = (b * b^{-1}) * a = e * a = a = b * x$.

So: $b^{-1} * b * (b^{-1} * a) = b^{-1} * b * x$

i.e. $e * (b^{-1} * a) = e * x$

i.e. $b^{-1} * a = x$.

The other result follows similarly. Q.E.D.

Note: these last two properties show how careful we must be in the general case when the group is not (necessarily) commutative. For example: $(a * b)^{-1} \neq a^{-1} * b^{-1}$. If the group is commutative then, everything is simpler:

$$(a * b)^{-1} = a^{-1} * b^{-1} \quad \text{and} \quad a * b^{-1} = b^{-1} * a.$$

The second, for a *commutative* group under $+$, provides the definition of a unique difference: $a-b=a+(-b)=(-b)+a$. Similarly, for a *commutative* group under $\times$, it gives the definition of a unique quotient: $\dfrac{a}{b}=a\times\dfrac{1}{b}=\dfrac{1}{b}\times a$.

(iii) *Rings.* As a first development on the basis of a group, consider sets of double composition in which two binary operations are defined: sums and products. A general case of such sets of double composition is the ring:

DEFINITION: *A set* $R=\{a, b, c, \ldots\}$ *is a* **ring** *if the two operations* ($+$ *and* $\times$) *are such that:*

(1) *the elements form a commutative group* ($R+$) *under addition*

(2) *the non-zero elements as a set* ($R\times$) *under multiplication satisfy:*

$$\textit{Closure: } ab\in R \qquad \textit{Associative: } a(bc)=(ab)c$$

(3) *addition and multiplication are connected by the distributive rules:*

$$a(b+c)=ab+ac \quad \textit{and} \quad (a+b)c=ac+bc.$$

The implications are that R is fully obedient under addition, that multiplication has only a minimum of properties (closure and associative) and that multiplication is distributive over addition. Other properties (commutative, unity and reciprocals) of products are *not* assumed. It is not even assumed that cancellation (if $ab=ac$, for $a\neq 0$, then $b=c$) is a valid procedure. A ring may have zero divisors, i.e. non-zero a and b such that $ab=0$. This leaves it open to specify more particular, i.e. more specialised, kinds of rings. For example, there is the commutative ring, as opposed to the non-commutative type, for which $ab=ba$. Increasing the specialisation by adding further multiplicative properties, we get to a penultimate stage with the *integral domain*: a commutative ring with identity (unity) for which cancellation holds for products. All that this kind of ring lacks is a reciprocal for each element. The set J of integers is an example. The last, or completely obedient, stage is reached with a *field*: a commutative ring with identity (unity) and reciprocals.

(iv) *Fields.* The end of the development of sets of double composition (under $+$ and $\times$) is the field, a set which comprehends in itself two commutative groups, one for sums and the other for products, linked by the distributive rule. A direct and economical definition of a field can be given as an alternative:

DEFINITION: *A set $F=\{a, b, c, \dots\}$ is a* **field** *if it has at least two distinct elements and if operations of $+$ and $\times$ are defined so that:*

(1) *they are closed, associative and commutative, and connected by the distributive rule: $a(b+c)=ab+ac$*

(2) *there is always a solution of the equations $(x \in F)$:*

$$x+a=b \quad \text{and} \quad ax=b \quad \text{for any } a \text{ and } b \in F$$

except that $a \neq 0$ must be imposed for the second equation.

The requirement of two distinct elements ensures that there is something to act as zero for addition, and something different for unity in multiplication. Conditions (2) ensure that linear equations can be solved within a field. In effect, $x+a=b$ gives the difference $b-a$ and $ax=b$ gives the quotient $\frac{a}{b}$.

From the definition, it is easily established that the set of all elements of F is a commutative group under $+$ and that the set of all non-zero elements is a commutative group under $\times$, i.e. that F is the most specialised of rings. Checking back at the economical definition of a group in (ii) above, we find that all we need to prove is that e exists for $e * a=a$ and that a^{-1} exists for $a^{-1} * a=e$, both when $*$ is $+$ and when $*$ is $\times$. All the rest follows, e.g. that e is unique in both cases (0 for sums and 1 for products). The proof for multiplication is as follows: There is always a solution of $ax=b$, i.e. for $xa=b$ (commutative property). Define e as a solution of $xb=b$, for some given b: $eb=b$. Now let a be any member of F. Then x exists for $bx=a$ and so: $ea=ebx=bx=a$. Hence e is such that $ea=a$ for any a, i.e. e is the identity required. Further, consider the equation $xa=e$. Write the solution $x=a^{-1}$, so that $a^{-1}a=e$, and a^{-1} is the reciprocal of a required.

(v) *Vector Spaces.* The double composition of a ring or field is achieved by supplementing an additive group with the second operation of multiplication. An additive group can be made into a system of double composition in a different way, the second operation being a new one: scalar multiplication. The result is a vector space V, comprising a set $\{u, v, w, \dots\}$ of elements called *vectors*. To write scalar products of vectors in V, we need another set, i.e. a set of *scalars*. This is an outside set $\{a, b, c, \dots\}$, different in nature from V. We take it to be a field and denote it by F. *Scalar multiplication* is then an operation which takes a vector u from the main set V and a scalar a from the outside field F, and which gives the product au as a vector of V.

DEFINITION: *$V=\{u, v, w, \dots\}$ is a* **vector space** *over the field $F=\{a, b, c, \dots\}$ of scalars if the two operations of addition of vectors and of the product of a vector by a scalar are such that:*

(1) *V is a commutative group under addition*

(2) *scalar multiplication is closed ($au \in V$) with properties:*

Associative: $a(bu)=(ab)u$

Distributive: $a(u+v)=au+av$

and $(a+b)u=au+bu$

and the unit scalar (1) *is such that* $1u=u$.

Notice that a vector space does not need the operation of multiplication of vectors, just as a ring of field does not need scalar products. Each system has two operations: sums and scalar products in one case, sums and products in the other. The following is the simplest example of a vector space.

Write the vector $v=(x, y)$, a pair of real numbers. Define addition:

$$v_1+v_2=(x_1, y_1)+(x_2, y_2)=(x_1+x_2, y_1+y_2).$$

Call the field of real numbers F, the set of scalars used. Define scalar products: $av=a(x, y)=(ax, ay)$. Then the set of all number pairs v is a vector space V over F. The vectors may be shown as points P or vectors OP in a plane, or as complex numbers on an Argand diagram. In any case, the difference between V and F is clear: V has elements which are *number pairs* (e.g. points in a plane) and F comprises *real numbers*. The example is easily generalised to a set of n-tuples of numbers.

15.4. Limits and continuity. (Reference: Chapter 9.) A general definition of a limit process, applicable in all cases, is given first, followed by particular applications of it.

(i) *Limit Processes.* A limit process depends on the initial specification of a set of stages through which the process is to run.

DEFINITION: *A set of* **stages** *$\{M, N, P, \dots\}$ is a set with the property that* **some** *stages are related $M>N$ (meaning M more advanced than, or contained in, N) where the relation $>$ satisfies the conditions:*

(i) **Transitivity:** *if $M>N$ and $N>P$, then $M>P$*

(ii) **Extension:** *if M and N are* **any** *stages, then there is a stage P such that $P>M$ and $P>N$.*

The relation $>$ is akin to an ordering (see the properties of order in 2.2 and 7.8) but the stages are *not* necessarily completely ordered, i.e. the stages are *not* necessarily contained one within the other as a set of nesting intervals. The import of property (ii) is that stages M and N can overlap, in which case there is another (and more advanced) stage P in the overlap. A typical set of stages is a set of neighbourhoods N of a fixed point $x=\alpha$ in an interval $[a, b]$. The neighbourhoods can overlap in all kinds of ways but a sequence of nesting intervals can be picked out, each more advanced than (contained in) another.

Definition: *A* **limit process** *exists if there is a mapping of a set of stages N onto a set of intervals $F(N)$ such that $F(M)$ is contained in $F(N)$ whenever $M>N$. The* **final residue** *of the limit process is the interval F which is the intersection of all $F(N)$. The limit process is* **convergent** *to L if F consists of the single element L: $\underset{N}{Lim}\, F(N)=L$.*

The intervals referred to are closed, i.e. $[a, b]$ is the set of x such that $a\leqslant x\leqslant b$. The definition states that F is an interval; this needs to be proved.

Proof: Any two intervals $F(M)$ and $F(N)$ must overlap and contain a third $F(P)$. For, there is a stage P such that $P>M$, implying $F(P)$ contained in $F(M)$, and such that $P>N$, implying $F(P)$ contained in $F(N)$. Let $F(N)=[c_n, d_n]$ overlap with $F(M)=[c_m, d_m]$. Then $c_n<d_m$ so that the set of c_n (all n) has an upper bound and so a LUB c. Similarly the set of d_n (all n) has a GLB d. Since $c_n\leqslant d_n$ all n, $c\leqslant d$ defining an interval $[c, d]$. This may be a single element (case $c=d$), or a finite interval (case $c<d$). Now, if λ is in $[c, d]$, then $c_n\leqslant c\leqslant\lambda\leqslant d\leqslant d_n$ all n, i.e. λ is in all $F(N)$. Conversely, if λ is in all $F(N)$, then $c_n\leqslant\lambda\leqslant d_n$ all n, i.e. λ is an upper bound of c_n and $c\leqslant\lambda$. Similarly $\lambda\leqslant d$. Hence λ is in $[c, d]$. So F is the interval $[c, d]$.

Q.E.D

There are various equivalent expressions of convergence:

Theorem: *$F(N)$ converges to L if and only if:*

(1) *for a given positive ϵ (however small), there is a stage N so that $L-\epsilon\leqslant c_n\leqslant L+\epsilon$ for all c_n of $F(N)$*

(2) *for a given $L' \neq L$ (however close), there is a stage N so that $F(N)$ does not contain L'*

(3) *for a given neighbourhood N_L of L (however small), there is a stage N so that $F(N)$ is contained in N_L.*

Proof: Only (1) need be established; the others follow from it. Directly: suppose $F(N)$ converges to L and take any positive ϵ. Then there is a stage M such that $L-\epsilon$ is not in $F(M)$ and a stage P such that $L+\epsilon$ is not in $F(P)$. Take N more advanced than M and P. $N > M$ implies $F(N)$ contained in $F(M)$ and $N > P$ implies $F(M)$ contained in $F(P)$. Hence N is the required stage so that $F(N)$ excludes both $L-\epsilon$ and $L+\epsilon$. Conversely: suppose N exists for $F(N)$ to exclude $L-\epsilon$ and $L+\epsilon$, for a given ϵ. Take $L' \neq L$ and write $\epsilon = | L - L' |$. For this ϵ, let N be the stage so that $F(N)$ excludes $L \pm \epsilon$, i.e. excludes L'. Hence the final residue F excludes L', which is *any* element $\neq L$. So F contains L only and

$$\operatorname*{Lim}_{N} F(N) = L. \qquad \text{Q.E.D.}$$

(ii) *Limits of a Function of a Real Variable.* Consider $\operatorname*{Lim}_{x \to a} f(x)$ for a function $f(x)$ defined on the domain X, where X need not include the given point $x = \alpha$ but must include some points in each neighbourhood of α. Take as stages N the set of neighbourhoods of α, satisfying the required conditions for a set of stages. If $f(x)$ is bounded in each N, there is a *smallest* interval $F(N)$ containing all $f(x)$ for $x \in N$. There is a mapping of N onto $F(N)$ and, by the definition of $F(N)$, it follows that $F(M)$ is contained in $F(N)$ wherever $M > N$ (M contained in N). Hence, if $f(x)$ is bounded in each N, there is a limit process over stages N. Then:

DEFINITION: *If $f(x)$ is bounded and the limit process $F(N)$ converges to L over the stages N, then $\operatorname*{Lim}_{x \to a} f(x) = L$. Necessary and sufficient conditions are that, for any positive ϵ (however small), there is a neighbourhood $[a_n, b_n]$ of α so that*

$$L - \epsilon \leqslant f(x) \leqslant L + \epsilon \quad \textit{for all } x \textit{ in } a_n \leqslant x \leqslant b_n.$$

The stated conditions are those of the Theorem of (i). The following properties follow. If $f(x)$ and $g(x)$ have limits as $x \to \alpha$, then:

$$\operatorname*{Lim}_{x \to a} \{f(x) \pm g(x)\} = \operatorname*{Lim}_{x \to a} f(x) \pm \operatorname*{Lim}_{x \to a} g(x);$$

$$\operatorname*{Lim}_{x\to a}\{f(x)\times g(x)\}=\operatorname*{Lim}_{x\to a}f(x)\times\operatorname*{Lim}_{x\to a}g(x);$$

$$\operatorname*{Lim}_{x\to a}\frac{f(x)}{g(x)}=\operatorname*{Lim}_{x\to a}f(x)\Big/\operatorname*{Lim}_{x\to a}g(x)\quad\text{if}\quad\operatorname*{Lim}_{x\to a}g(x)\neq 0.$$

Proof: take the case of a sum to illustrate. Let N' be stages for $f(x)\to L'$ as $x\to\alpha$ i.e. the smallest intervals $F(N')=[c_n', d_n']$ converge to L'. Let N'' be stages for $g(x)\to L''$ as $x\to\alpha$, i.e. the smallest intervals $G(N'')=[c_n'', d_n'']$ converge to L''. Take N as the stages common to N' and N''. Since $f(x)$ and $g(x)$ are bounded in each N, so is $f(x)+g(x)$, and there is a least interval $H(N)$ containing all $f(x)+g(x)$ in N. But $H(N)$ is contained in $[c_n'+c_n'', d_n'+d_n'']$ which converges to $L'+L''$. So $H(N)$ converges to $L'+L''$ and this is the limit of $f(x)+g(x)$ as $x\to\alpha$. Q.E.D.

The same definition and properties apply to $\operatorname*{Lim}_{x\to\infty}f(x)=L$. It is only necessary to re-specify the stages N so that N is the set of all $x\geqslant x_n$ for a given x_n (however large). $\operatorname*{Lim}_{x\to-\infty}f(x)$ follows similarly for stages N, where N is the set of all $x\leqslant x_n$ for a given x_n which is negative (however large).

(iii) *Continuity of a Function of a Real Variable.* Consider a function $f(x)$ defined on the domain X and let $x=\alpha$ be particular point of X.

DEFINITION: *$f(x)$ is* **continuous** *at $x=\alpha$ if* (1) *$\operatorname*{Lim}_{x\to a}f(x)$ exists,* (2) *$f(\alpha)$ is defined, and* (3) *$\operatorname*{Lim}_{x\to a}f(x)=f(\alpha)$.*

It follows from the properties of limits in (ii):

> If $f(x)$ and $g(x)$ are continuous at $x=\alpha$, then so are $f(x)\pm g(x)$, $f(x)\times g(x)$, and $f(x)/g(x)$, provided that $g(\alpha)\neq 0$ in the last case.

It is important to distinguish between the case where $f(x)$ is continuous at one value $x=\alpha$ and that where $f(x)$ is continuous for all $x\in X$. If $f(x)$ is continuous at $x=\alpha$, then it is locally bounded, i.e. bounded over some neighbourhood N of α. If $f(x)$ is continuous over X, then it is locally bounded everywhere in X. The most important result arises when $f(x)$ is defined and continuous over an *interval*, as a particular case of the domain X:

THEOREM: *If $f(x)$ is continuous over the interval $a\leqslant x\leqslant b$, then:*

(1) *$f(x)$ is bounded in the interval*

(2) *the range of $f(x)$ is itself an interval: $c\leqslant f(x)\leqslant d$*

(3) *$f(x)$ attains its GLB $c=f(\alpha)$ at some α ($a\leqslant\alpha\leqslant b$) and attains its LUB $d=f(\beta)$ at some β ($a\leqslant\beta\leqslant b$).*

Proof: (1) Let S be the set of λ such that $f(x)$ has an upper bound in $[a, \lambda]$, where $a\leqslant\lambda\leqslant b$. S is not empty ($\lambda=a$ at least) and it has an upper bound $\lambda\leqslant b$. Hence S has a LUB, say μ. But $f(x)$ is continuous at μ ($a\leqslant\mu\leqslant b$) and there is a neighbourhood of μ in which $f(x)$ is bounded. Hence $f(x)$ has an upper bound in the combined interval, which can be written $[a, k]$, made up of the intersection of $[a, \mu]$ and the neighbourhood of μ. Since $k>\mu$ (LUB of S), this is impossible except when k is *outside* the interval $[a, b]$. So, $f(x)$ has an upper bound in $[a, b]\subset[a, k]$. Similarly $f(x)$ has a lower bound in $[a, b]$.

Q.E.D.

(2) Write c as the GLB and d as the LUB of $f(x)$ over the interval $[a, b]$. We have to prove that, if β is any value in the interval $c\leqslant\beta\leqslant d$, then there is an $\alpha\,(a\leqslant\alpha\leqslant b)$ such that $f(\alpha)=\beta$. If $f(a)=\beta$, there is nothing to prove. Otherwise, suppose $f(a)<\beta$. (A similar argument holds if $f(a)>\beta$.) Consider the set S of λ such that $f(x)$ has an upper bound $<\beta$ over $[a, \lambda]$, where $a\leqslant\lambda\leqslant b$. Let α be the LUB of λ in S and suppose $f(\alpha)>\beta$. Then continuity of $f(x)$ at α gives a neighbourhood of α with $f(x)>\beta$ and this neighbourhood must overlap with S where $f(x)<\beta$. This is a contradiction, so that $f(\alpha)\not>\beta$. Now suppose $f(\alpha)<\beta$, so that there is a neighbourhood of α also with $f(x)<\beta$. Form the combined interval $[a, k]$ of the interval $[a, \alpha]$ and the neighbourhood of α, so that $f(x)$ has an upper bound $<\beta$ over $[a, k]$. Again, as in the proof of (1), k must lie outside $[a, b]$. Hence $f(x)$ has an upper bound $<\beta$ in $[a, b]$, contained in $[a, k]$. This contradicts the facts that the LUB of $f(x)$ is d and that β is set so that $\beta\leqslant d$. Hence $f(\alpha)\not<\beta$. The only possibility is $f(\alpha)=\beta$.

Q.E.D.

(3) follows at once from the proof of (2).

Consequently, if $f(x)$ is continuous over an interval, then both the domain and the range of the function are intervals. The result (3) is often called *Weierstrass' Maximum Theorem* after Weierstrass (1815–97). It ensures that $f(x)$ attains its LUB d at some point of the interval ($a\leqslant x\leqslant b$) and this LUB might be called the *maximum* of $f(x)$ in the interval. However, it is customary to limit the term to a *local* maximum (11.2). The LUB is better termed a *supremum* of $f(x)$ in the interval; it is a maximum of maxima (of which there may be several). Hence, Weierstrass' Theorem states that $\underset{x}{\operatorname{Sup}} f(x)=d$, attained at some point of the interval $(a \leqslant x \leqslant b)$. Similar remarks apply to the GLB c, the minimum of minima of $f(x)$ in the interval.

(iv) *Derivatives.* Consider a function $f(x)$ defined on the interval $[a, b]$ which contains $x=\alpha$ as an interior point. Then:

DEFINITION: *The* **derivative** *of $f(x)$ at $x=\alpha$ is:*

$$f'(\alpha)=\underset{x\to a}{Lim}\frac{f(x)-f(\alpha)}{x-\alpha} \quad \textit{if the limit exists.}$$

Write $F(x)=\dfrac{f(x)-f(\alpha)}{x-\alpha}$ so that $F(x)$ is *not* defined at $x=\alpha$. It is still valid, however, to seek the limit of $F(x)$ as $x\to\alpha$. If the limit exists, it is $f'(\alpha)$; if the limit does not exist, neither does $f'(\alpha)$. A *necessary condition* is:

THEOREM: *If $f'(\alpha)$ exists, then $f(x)$ is continuous at $x=\alpha$.*

Proof: Write $\phi(x)=F(x)(x-\alpha)+f(\alpha)=f(x)$ except at $x=\alpha$, where $F(x)$ is written above. Though $\phi(x)$ is not defined at $x=\alpha$, the limit does exist:

$$\phi(x)\to f'(\alpha)\times 0+f(\alpha)=f(\alpha)$$

as $x\to\alpha$. Here, since $\phi(x)=f(x)$ $(x\neq\alpha)$, $f(x)\to f(\alpha)$ as $x\to\alpha$, i.e. $f(x)$ is continuous at $x=\alpha$. Q.E.D.

15.5. Integrals: the fundamental theorem of the calculus. (Reference: Chapter 10.) The function $f(x)$ is defined on the interval $[a, b]$ where $a<b$, and it is bounded on the interval. A partition P of $[a, b]$ is any set of n segments (n any positive integer) with dividing points $a=x_0, x_1, x_2, \ldots x_n=b$. A refinement P' of P is another partition which includes $x_0, x_1, x_2, \ldots x_n$ among its dividing points. Write $P'>P$ when P' is a refinement of (more advanced than) P. Then the partitions P are *stages* of a limit process (15.4) since they have both the necessary properties of transitivity and extension. The first of these is obvious and for the second, if P_1 and P_2 are any two partitions, take P as including all the dividing points of both to ensure that $P>P_1$ and $P>P_2$.

Form the sum $\sum\limits_{r=1}^{n} f(x_r')(x_r-x_{r-1})$ over P, where x_r' $(r=1, 2, \ldots n)$

is any selected point in the rth segment $(x_{r-1} \leqslant x_r' \leqslant x_r)$. Since $f(x)$ is bounded, it has a GLB L_r and a LUB G_r in the rth segment:

$$L_r \leqslant f(x_r') \leqslant G_r.$$

So the sum is contained in the interval

$$\left[\sum_{r=1}^{n} L_r(x_r - x_{r-1}),\ \sum_{r=1}^{n} G_r(x_r - x_{r-1})\right]$$

and this is the smallest such interval. The interval depends on P but not on the choice of x_r'. Hence:

Definition: *The* **Riemann Sum** *for $f(x)$ and the partition P of $[a, b]$ is:*

$$F(P) = \left[\sum_{r=1}^{n} L_r(x_r - x_{r-1}),\ \sum_{r=1}^{n} G_r(x_r - x_{r-1})\right]$$

and this interval is denoted $\overset{b}{\underset{a}{S}} f(x)\Delta x$ over P.

The Riemann Sum is the *smallest* interval containing

$$\sum_{r=1}^{n} f(x_r')(x_r - x_{r-1})$$

for all choices of x_r' in a given partition P. It is named after Riemann (1826–66). Among the properties which follow immediately from the definition, note:

(1) If L is the GLB and G the LUB of $f(x)$ in $[a, b]$, then $\overset{b}{\underset{a}{S}} f(x)\Delta x$ over P is contained in the interval $[L(b-a), G(b-a)]$ for all P.

(2) $\overset{b}{\underset{a}{S}} f(x)\Delta x$ over $P = \overset{c}{\underset{a}{S}} f(x)\Delta x$ over $P_1 + \overset{b}{\underset{c}{S}} f(x)\Delta x$ over P_2 $(a < c < b)$ for two parts P_1 and P_2 of P with dividing point c.

The meaning of the second property is that the lower point (upper point) of the interval $\overset{b}{\underset{a}{S}}$ is the sum of the lower points (upper points) of the intervals $\overset{c}{\underset{a}{S}}$ and $\overset{b}{\underset{c}{S}}$.

From the definition, it also follows that, if $P' > P$, then the interval $F(P')$ is contained in the interval $F(P)$. Proof: suppose that the segment $(x_r - x_{r-1})$ of P is split by $\bar{x}$ into two segments of P': $(\bar{x} - x_{r-1})$ and $(x_r - \bar{x})$. If L_r' and L_r'' are the GLB of $f(x)$ in these two segments, then $L_r' \geqslant L_r$, $L_r'' \geqslant L_r$. The contribution of the segment to the lower end of $F(P)$ is $L_r(x_r - x_{r-1})$; the contribution to the lower end of $F(P')$ is $L_r'(\bar{x} - x_{r-1}) + L_r''(x_r - \bar{x}) \geqslant L_r(x_r - x_{r-1})$ i.e. the lower end of $F(P') \geqslant$ that of $F(P)$. Similarly, the upper end of $F(P') \leqslant$ that of $F(P)$. This extends to any subdivision of P. Q.E.D.

Hence there is a mapping of the stages P onto the intervals $F(P)$ such that $F(P')$ is contained in $F(P)$ whenever $P' > P$, i.e. a limit process (15.4)

Definition: *If $f(x)$ is bounded on $[a, b]$, where $a < b$, the* **Riemann Integral** *is the limit of $F(P)$ as P advances, if the limit exists:*

$$\int_a^b f(x)\,dx = \underset{P}{Lim} \overset{b}{\underset{a}{S}} f(x)\Delta x \quad over\ P.$$

If the limit process $F(P)$ does not converge, the integral does not exist. From the properties (1) and (2) of Riemann Sums, it follows:

(i) If L is the GLB and G the LUB of $f(x)$ in $[a, b]$, then:

$$L(b-a) \leqslant \int_a^b f(x)\,dx \leqslant G(b-a).$$

(ii) If $a < c < b$, then $\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$

In both cases, it is assumed that the integral of $f(x)$ exists.

It remains to establish the *Fundamental Theorem of the Calculus.* In its simplest, if not quite its strongest, form:

Theorem: *If $f(x)$ is continuous on the interval $[a, b]$, then at each x of $[a, b]$: $F(x) = \int_a^x f(u)\,du$ exists, is continuous and has derivative $F'(x) = f(x)$.*

Proof: Let x and $x+h$ be points in $[a, b]$. Since $f(x)$ is continuous, $f(x+h) \to f(x)$ as $h \to 0$, which can be expressed as follows, by (3) of the Theorem of 15.4 (i) above. Given a neighbourhood N of $f(x)$, there exists a neighbourhood of x for which the range of $f(x)$ is contained in N. Further, by the Theorem of 15.4 (iii) (Weierstrass' Maximum Theorem), $f(x)$ is bounded and attains its bounds in any interval within $[a, b]$. Let c be the GLB and d the LUB of $f(x)$ in the interval $[x, x+h]$. Then, given N, there is an h such that $[c, d]$ is contained in N.

Since $f(x)$ is bounded, Riemann Sums exist for a partition P of $[a, b]$, and for the constituent parts of P separately. Consider

$\overset{x}{\underset{a}{S}} f(x)\Delta x$, $\overset{x+h}{\underset{a}{S}} f(x)\Delta x$ and $\overset{x+h}{\underset{x}{S}} f(x)\Delta x$. From the limit process, $\overset{x}{\underset{a}{S}} f(x)\Delta x$ as an interval for various P has a final residue which (as shown in 15.4) is itself an interval. Write the final residue, not dependent on P, as $[L_a(x), G_a(x)]$ which indicates that it depends both on the fixed lower point a and on the variable upper point x. Similarly, the final residue of $\overset{x+h}{\underset{a}{S}} f(x)\Delta x$ over P is the interval $[L_a(x+h), G_a(x+h)]$, and that of $\overset{x+h}{\underset{x}{S}} f(x)\Delta x$ is $[L_x(x+h), G_x(x+h)]$.

The position, now, is that $[L_x(x+h), G_x(x+h)]$ as final residue is contained in the interval $\overset{x+h}{\underset{x}{S}} f(x)\Delta x$ over P, and this is contained in $[c(x+h-x), d(x+h-x)]=[ch, dh]$ by property (1) above. Hence:

$$\left[\frac{L_x(x+h)}{h}, \frac{G_x(x+h)}{h}\right]$$

is contained in $[c, d]$ i.e. in N (given). But, by property (2) above,

$$\overset{x+h}{\underset{x}{S}} f(x)\Delta x = \overset{x+h}{\underset{a}{S}} f(x)\Delta x - \overset{x}{\underset{a}{S}} f(x)\Delta x.$$

This means that

$$L_x(x+h)=L_a(x+h)-L_a(x) \quad \text{and} \quad G_x(x+h)=G_a(x+h)-G_a(x).$$

So: $$\left[\frac{L_a(x+h)-L_a(x)}{h}, \frac{G_a(x+h)-G_a(x)}{h}\right] \text{ is contained in } N.$$

This is true, given N, for any sufficiently small h (positive or negative).

A limit process is now set up. As N, any neighbourhood of $f(x)$, contracts on $f(x)$, h gets smaller and the function $L_a(x)$ has a derivative:

$$L_a'(x)=\underset{h\to 0}{\text{Lim}} \frac{L_a(x+h)-L_a(x)}{h}=f(x).$$

Similarly: $$G_a'(x)=\underset{h\to 0}{\text{Lim}} \frac{G_a(x+h)-G_a(x)}{h}=f(x).$$

Hence $G_a(x)-L_a(x)$ has zero derivative, i.e. $G_a(x)-L_a(x)=$constant, which must be zero since $G_a(a)=L_a(a)=0$. So $G_a(x)=L_a(x)$ and each has derivative $f(x)$. Write the common value $F(x)$, such that $F'(x)=f(x)$. But the final residue of $\overset{x}{\underset{a}{S}} f(x)\Delta x$ is $[L_a(x), G_a(x)]$, now shown to be a single value $F(x)$. Hence $\int_a^x f(x)\,dx$ exists; it is $F(x)$ where $F(x)$ is continuous with derivative $F'(x)=f(x)$. Q.E.D.

15.6. Absolute and uniform convergence. (Reference: Chapter 11.) An infinite series Σu_n is given, with no restriction on the sign of u_n.

DEFINITION: Σu_n *is* **absolutely convergent** *if* $\Sigma \mid u_n \mid$ *is convergent.* The following result, called *Dirichlet's Theorem* after Dirichlet (1805–59), is derived from the definition:

THEOREM: *The sum of an absolutely convergent series does not depend on the order taken for the terms of the series.*

Proof: given Σu_n, absolutely convergent, write Σv_n and Σw_n where:

$$v_n = u_n \ (u_n \text{ positive}) \text{ and } =0 \ (u_n \text{ negative})$$
$$w_n = 0 \ (u_n \text{ positive}) \text{ and} = -u_n \ (u_n \text{ negative})$$

so that Σv_n is the sum of the positive terms of Σu_n with zero terms interspersed and Σw_n is similarly defined for the negative terms of Σu_n. Then

$$\Sigma u_n = \Sigma v_n - \Sigma w_n \quad \text{and} \quad \Sigma \mid u_n \mid = \Sigma v_n + \Sigma w_n$$

and both Σv_n and Σw_n (series of non-negative terms) are convergent. Let $\Sigma v_n = S$. Then the sum of any (finite) number of terms of Σv_n is $\leqslant S$. Let v_n' be any re-arrangement of the order of terms in Σv_n. Then the sum of any (finite) number of terms of $\Sigma v_n'$ is $\leqslant S$, since all the v_n''s are v_n's. Hence $\Sigma v_n'$ is convergent to sum $S' \leqslant S$. Similarly, starting with $\Sigma v_n' = S'$ and re-arranging to $\Sigma v_n = S$, we have $S \leqslant S'$. Hence, $S = S'$. Similarly for Σw_n and any re-arrangement $\Sigma w_n'$. So, for any re-arrangement $\Sigma u_n'$ of Σu_n:

$$\Sigma u_n' = \Sigma v_n' - \Sigma w_n' = \Sigma v_n - \Sigma w_n = \Sigma u_n. \qquad \text{Q.E.D.}$$

The important application of this Theorem is to establish the validity of the process of *multiplication of series*, provided that they are absolutely convergent:

THEOREM: *If Σu_n and Σv_n are absolutely convergent, then the product series*

$$u_1v_1 + (u_1v_2 + u_2v_1) + (u_1v_3 + u_2v_2 + u_3v_1) + \ldots$$

is absolutely convergent to sum $\Sigma u_n \times \Sigma v_n$.

Proof: all possible pairings of terms can be set out in double array:

$$\begin{array}{cccc} u_1v_1 & u_1v_2 & u_1v_3 & \dots \\ u_2v_1 & u_2v_2 & u_2v_3 & \dots \\ u_3v_1 & u_3v_2 & u_3v_3 & \dots \\ \cdot & \cdot & \cdot & \cdot \end{array}$$

Put the whole set into a single series in two ways:

(i) $u_1v_1 + (u_1v_2 + u_2v_1) + (u_1v_3 + u_2v_2 + u_3v_1) + \dots$

where the nth group is composed of terms in a diagonal of the array, starting from u_1v_n in the first row and running down from right to left. This is the product series.

(ii) $u_1v_1 + (u_1v_2 + u_2v_2 + u_2v_1) + (u_1v_3 + u_2v_3 + u_3v_3 + u_3v_2 + u_3v_1) + \dots$

where the nth group is composed of terms u_nv_n, *plus* all terms above this in the nth column of the array, *plus* all terms to its left in the nth row of the array. Then:

$$\begin{array}{ll} \text{first group} & = u_1v_1 \\ \text{second group} & = (u_1+u_2)(v_1+v_2) - u_1v_1 \\ \text{third group} & = (u_1+u_2+u_3)(v_1+v_2+v_3) - (u_1+u_2)(v_1+v_2) \\ \cdot \quad \cdot \quad \cdot & \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \end{array}$$

$$\text{Sum of } n \text{ groups} = (u_1+u_2+\dots+u_n)(v_1+v_2+\dots+v_n)$$
$$\to \Sigma u_n \times \Sigma v_n \quad \text{as } n \to \infty.$$

Hence the series (ii) is absolutely convergent to $\Sigma u_n \times \Sigma v_n$. By Dirichlet's Theorem, (i) as a re-arrangement of (ii) is absolutely convergent to $\Sigma u_n \times \Sigma v_n$. Q.E.D.

Consider a series $\Sigma u_n(x)$, each term a function of x defined on some domain X. A necessary and sufficient condition that $\Sigma u_n(x)$ is absolutely convergent, i.e. $\Sigma \mid u_n(x) \mid$ convergent, at each x of X is:

Given ϵ, there is an integer $N(x)$ for each x of X such that

$$\sum_{s=n+1}^{m} \mid u_s(x) \mid \leqslant \epsilon \quad \text{for all } m > n > N(x) \quad \dots\dots\dots\dots\dots\dots (1)$$

This follows from the Theorem of 11.3 since the sum shown, from the $(n+1)$th to the mth term inclusive, is the difference between the sums to n and to m terms of the series $\Sigma \mid u_n(x) \mid$ of positive terms. In general, the choice of N in (1) must depend on the x taken; it cannot be assumed that the same N will serve for all x of X. If this happens to be so, the series is uniformly convergent as well as absolutely convergent:

DEFINITION: *The series $\Sigma u_n(x)$ is* **uniformly convergent** *over X if it is absolutely convergent so that* (1) *holds for a* **constant** N.

The following is one of the tests for uniform convergence:

M Test: The series $\Sigma u_n(x)$ is uniformly convergent over X if a convergent series ΣM_n of positive and constant terms exists so that $|u_n(x)| \leqslant M_n$ for each n and for all x of X.

Proof: Since ΣM_n converges, given ϵ there is an integer N so that $\sum_{s=n+1}^{m} M_s \leqslant \epsilon$ for all $m>n>N$. Hence,

$$\sum_{s=n+1}^{m} |u_s(x)| \leqslant \sum_{s=n+1}^{m} M_s \leqslant \epsilon \text{ for all } m>n>N.$$

Here N is not dependent on x of X, i.e. $\Sigma u_n(x)$ is uniformly convergent. Q.E.D.

Apply these results to a power series. By (1), $\Sigma a_n x^n$ has radius of convergence r (absolutely convergent over $-r<x<r$) if and only if:

Given ϵ, then there is an integer $N(x)$ for each x $(-r<x<r)$ such that

$$\sum_{s=n+1}^{m} |a_s|\,|x|^s \leqslant \epsilon \quad \text{for all } m>n>N(x) \qquad \ldots\ldots\ldots\ldots\ldots(2)$$

For uniform convergence, $N(x)$ in (2) needs to be replaced by constant N.

THEOREM: *If $\Sigma a_n x^n$ has radius of convergence r, the series is uniformly convergent over $-h \leqslant x \leqslant h$ for any constant $h<r$.*

Proof: Since $\Sigma a_n x^n$ is absolutely convergent over $-r<x<r$, the series is absolutely convergent for $x=h$. The series of positive constant terms $\Sigma\,|a_n|\,h^n$ is convergent. But $|a_n x^n| \leqslant |a_n|\,h^n$ for all x in $-h \leqslant x \leqslant h$. Hence, by the M Test, $\Sigma a_n x^n$ is uniformly convergent over $-h \leqslant x \leqslant h$. Q.E.D.

The significance of this result is that, within the radius of convergence, a power series is uniformly convergent; this implies that it is absolutely convergent and, hence, convergent — though not conversely. If we write the sum $f(x)=\sum_{s=0}^{n} a_n x^n + R_n(x)$, then $R_n(x) \rightarrow 0$

as $n \to \infty$ within the radius of convergence. So, given ϵ, there is an integer N such that, for each x within the radius of convergence:

$$R_n(x) \leqslant \epsilon \quad \text{for all } n > N \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

For absolute convergence, (3) may hold only for $N(x)$ depending on x; for uniform convergence, it holds for constant N. The theorem shows that (3) holds for constant N *within* the radius of convergence of the power series.

15.7. Exponential and logarithmic functions. (Reference: Chapter 12.) The logarithmic function, its inverse the exponential function, and the extension to a power function, can all be derived from a single integral, that of the algebraic function $y = 1/x$. In the domain $x > 0$, the function $y = 1/x$ is continuous and decreasing. Hence, its integral exists in this domain.

DEFINITION: *log* $x = \int_1^x \frac{du}{u}$ *for* $x > 0$.

Two basic properties follow:

(1) $\log x$ is continuous and increasing, with derivative

$$D \log x = \frac{1}{x} > 0 \ (x > 0)$$

(2) $\log (xy) = \log x + \log y$; $\log \left(\frac{1}{x}\right) = -\log x$; $\log \left(\frac{x}{y}\right) = \log x - \log y$; $\log (x^n) = n \log x$ for integral n.

Proof: (1) is a consequence of the Fundamental Theorem of the Calculus. The result for $\log (xy)$ in (2) is established by substituting $u = yt$ (for given y) in the definition:

$$\int \frac{du}{u} = \int \frac{1}{yt}\, y\, dt = \int \frac{dt}{t}.$$

As t ranges from 1 to x, u ranges from y to xy:

$$\int_y^{xy} \frac{du}{u} = \int_1^x \frac{dt}{t} \quad \text{i.e.} \int_1^{xy} \frac{du}{u} = \int_1^x \frac{dt}{t} + \int_1^y \frac{du}{u}.$$

Hence $$\log xy = \log x + \log y.$$

As a particular case: $\log \left(x \frac{1}{x}\right) = \log x + \log \left(\frac{1}{x}\right)$.

But $$\log\left(x\frac{1}{x}\right)=\log 1=0 \quad \text{by the definition.}$$

Hence $$\log\left(\frac{1}{x}\right)=-\log x.$$

Then: $$\log\left(\frac{x}{y}\right)=\log\left(x\frac{1}{y}\right)=\log x+\log\left(\frac{1}{y}\right)=\log x-\log y.$$

Finally: $$\log x^2=\log (x\times x)=\log x+\log x=2\log x$$

and $$\log x^n=n\log x \quad \text{by induction.} \qquad \text{Q.E.D.}$$

Since the derivative of $\log x$ is everywhere positive $(x>0)$, $y=\log x$ is an increasing function. Further $\log 1=0$, so that $\log x<0$ for $0<x<1$, $\log x=0$ for $x=1$, and $\log x>0$ for $x>1$. Moreover, since $\frac{1}{x}$ decreases, $D\log x=\frac{1}{x}$ also decreases. The curve $y=\log x$ is of the form shown in Fig. 12.3 above, the slope of the tangent falling steadily as x increases. Finally:

DEFINITION: *the constant e is such that* $\int_1^e \frac{du}{u}=1.$

This implies that $\log e=1$.

A continuous and increasing function has an inverse which is also continuous and increasing. Hence:

DEFINITION: $y=\exp x$ *if* $x=\log y$ *for all* x *and for* $y>0$.

Two basic properties follow:

(i) $\exp x$ is continuous with derivative $D\exp x=\exp x>0$ (all x)

(ii) $\exp(x+y)=\exp x\times\exp y$; $\exp(-x)=\frac{1}{\exp x}$;

$\exp(x-y)=\frac{\exp x}{\exp y}$; $(\exp x)^n=\exp nx$ for integral n.

Proof: the inverse rule for derivatives establishes (i):

$$D_x\exp x=\frac{1}{D_y\log y}=\frac{1}{1/y}=y=\exp x.$$

Properties (ii) follow from the corresponding properties (2) of the logarithmic function.

The curve $y=\exp x$ is the same as that for $y=\log x$, with axes interchanged. It is of the form shown in Fig. 12.3 above. Hence

$y = \exp x$ is an increasing function, the tangent having increasing slope; $y < 1$ for $x < 0$, $y = 1$ at $x = 0$, and $y > 1$ for $x > 0$.

To relate to the constant e, such that $\log e = 1$, we have $\exp 1 = e$. Hence:

$$e^n = (\exp 1)^n = \exp n \quad \text{for integral } n, \text{ by (ii).}$$

This can be extended to:

$$e^r = \exp r \quad \text{for rational } r.$$

The rest is a matter of notation:

NOTATION: *the power* $e^x = \exp x$ *for all real* x.

If x is rational, then e^x is the power of e in the sense of elementary algebra; if x is not rational, e^x is simply taken as $\exp x$. The properties (ii) are then:

$$e^{x+y} = e^x e^y; \; e^{-x} = \frac{1}{e^x}; \; e^{x-y} = \frac{e^x}{e^y}; \; (e^x)^y = e^{xy}.$$

The expansion of the exponential function follows from Taylor's series:

$$e^x = \exp x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots + \frac{x^n}{n!} + \ldots$$

which is absolutely and uniformly convergent for all x. The expansion of the logarithmic function $\log(1+x)$ is derived from the geometric series by integration term by term:

$$\frac{1}{1+x} = 1 - x + x^2 - x^3 + \ldots + (-1)^{n-1} x^{n-1} + \ldots$$

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots + (-1)^{n-1} \frac{x^n}{n} + \ldots$$

These are absolutely and uniformly convergent for $|x| < 1$.

As a particular case of the exponential expansion for $x = 1$:

$$e = 1 + 1 + \frac{1}{2!} + \frac{1}{3!} + \ldots + \frac{1}{n!} + \ldots$$

The rational approximation to $e = 2{\cdot}71828\ldots$ is derived from this series, by taking a sufficient number of terms.

15.8. Circular functions. (Reference: Chapter 12.) Another single integral provides all the circular functions. It is again the integral of a simple algebraic function $y=\frac{1}{1+x^2}$, continuous and increasing for $x<0$, decreasing for $x>0$.

DEFINITION: *$\tan^{-1} x=\int_0^x \frac{du}{1+u^2}$ for all x.*

The basic properties are:

(1) $\tan^{-1} x$ is continuous and increasing, with $D \tan^{-1} x=\frac{1}{1+x^2}>0$ all x.

(2) $\tan^{-1} x+\tan^{-1} y=\tan^{-1}\frac{x+y}{1-xy}$.

Proof: (1) follows from the Fundamental Theorem of the Calculus. To establish (2), write $u=\frac{t-y}{1+yt}$ or $t=\frac{u+y}{1-yu}$ in the definition. Here, as u runs from 0 to x, t runs from y to z where $z=\frac{x+y}{1-xy}$. So:

$$\tan^{-1} x=\int_0^x \frac{du}{1+u^2}=\int_y^z \frac{1}{1+u^2}\, u'\, dt$$

where $$u=\frac{t-y}{1+yt};\ u'=\frac{(1+yt)-y(t-y)}{(1+yt)^2}=\frac{1+y^2}{(1+yt)^2}$$

and $$1+u^2=1+\frac{(t-y)^2}{(1+yt)^2}=\frac{(1+t^2)(1+y^2)}{(1+yt)^2}.$$

Hence: $$\tan^{-1} x=\int_y^z \frac{(1+yt)^2}{(1+t^2)(1+y^2)}\frac{1+y^2}{(1+yt)^2}\,dt=\int_y^z \frac{dt}{1+t^2}$$
$$=\int_0^z \frac{dt}{1+t^2}-\int_0^y \frac{dt}{1+t^2}=\tan^{-1} z-\tan^{-1} y.$$

So: $$\tan^{-1} x+\tan^{-1} y=\tan^{-1} z=\tan^{-1}\frac{x+y}{1-xy}.$$ Q.E.D.

DEFINITION: *the constant π is such that $\frac{1}{4}\pi=\int_0^1 \frac{du}{1+u^2}$.*

This implies that $\tan^{-1} 1=\frac{1}{4}\pi$. The range of $y=\tan^{-1} x$ is still to be found. It comes by substituting $u=\frac{1}{t}$ in the definition, t going from 1 to 0 as u increases from 1:

$$\int_1^\infty \frac{du}{1+u^2}=\int_1^0 \frac{1}{1+1/t^2}\left(-\frac{1}{t^2}\right)dt=\int_0^1 \frac{dt}{1+t^2}=\int_0^1 \frac{du}{1+u^2}$$

So: $$\int_0^\infty \frac{du}{1+u^2}=\int_0^1 \frac{du}{1+u^2}+\int_1^\infty \frac{du}{1+u^2}=2\int_0^1 \frac{du}{1+u^2}=\tfrac{1}{2}\pi.$$

Hence, $y=\tan^{-1} x \to \frac{1}{2}\pi$ as $x\to\infty$. Similarly, $y\to -\frac{1}{2}\pi$ as $x\to-\infty$. From the definition, property (1) and these results, it follows that $y=\tan^{-1} x$ is an increasing function, that $\tan^{-1} x<0$ for $x<0$, $\tan^{-1} x=0$ at $x=0$ and $\tan^{-1} x>0$ for $x>0$, and that $\tan^{-1} x\to\pm\frac{1}{2}\pi$ as $x\to\pm\infty$. The curve $y=\tan^{-1} x$ is of the form shown in Fig. 15.8*a*.

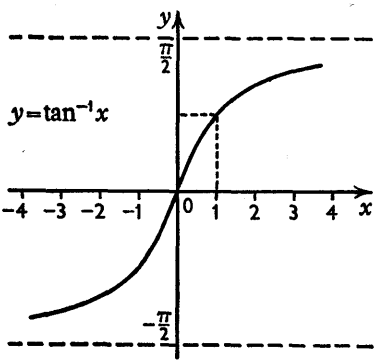


FIG. 15.8*a*

A continuous and increasing function has an inverse which is also continuous and increasing. Hence:

DEFINITION: *$y=\tan x$ if $x=\tan^{-1}y$ for x in the open interval $-\frac{1}{2}\pi<x<\frac{1}{2}\pi$.*

The range of $y=\tan x$ is all y, and $y\to\pm\infty$ as $x\to\pm\frac{1}{2}\pi$.

The basic properties are:

(i) $\tan x$ is continuous and increasing, with $D\tan x=1+\tan^2 x>0\left(-\frac{1}{2}\pi<x<\frac{1}{2}\pi\right)$

(ii) *Addition Formula:*

$$\tan(x+y)=\frac{\tan x+\tan y}{1-\tan x\tan y}$$

Proof: the inverse function rule gives (i)

$$D_x \tan x=\frac{1}{D_y \tan^{-1} y}=1+y^2=1+\tan^2 x.$$

(ii) follows from the corresponding result (2) for $\tan^{-1} x$.

The curve $y=\tan x$ on the domain $-\frac{1}{2}\pi<x<\frac{1}{2}\pi$ is that of $y=\tan^{-1} x$ with axes interchanged. It is continuous and increasing, as shown in Fig. 15.8*b*.

Two other circular functions follow:

DEFINITION: *$\cos x=\dfrac{1}{\sqrt{1+t^2}}$ and $\sin x=\dfrac{t}{\sqrt{1+t^2}}$ where*

$$t=\tan x\left(-\tfrac{1}{2}\pi<x<\tfrac{1}{2}\pi\right).$$

The basic properties are:

(*a*) $\tan x=\dfrac{\sin x}{\cos x}$; $\sin^2 x+\cos^2 x=1$

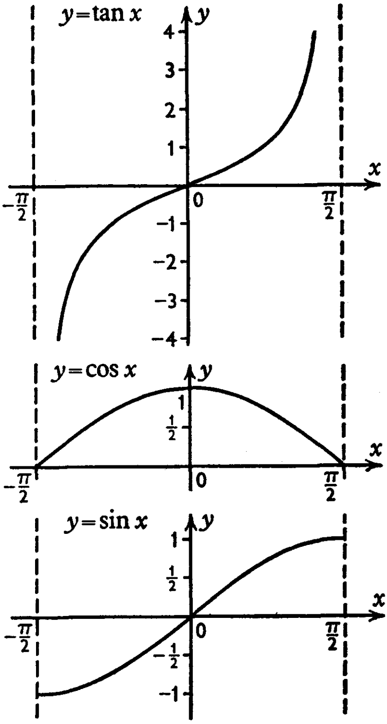


FIG. 15.8*b*

(*b*) cos x increases $(-\frac{1}{2}\pi<x<0)$ and then decreases $(0<x<\frac{1}{2}\pi)$ with derivative

$$D\cos x=-\sin x;$$

sin x increases $(-\frac{1}{2}\pi<x<\frac{1}{2}\pi)$ and $D\sin x=\cos x$.

(*c*) *Addition Formulae:*

$$\cos(x+y)=\cos x\cos y-\sin x\sin y$$
$$\sin(x+y)=\sin x\cos y+\cos x\sin y.$$

Proof: (*a*) follows from the definition. The derivatives of (*b*) are obtained by the function of a function rule with $t=\tan x$ and $D_xt=1+\tan^2x=1+t^2$:

$$D_x\cos x=D_t\frac{1}{\sqrt{1+t^2}}D_xt$$
$$=\left\{-\frac{2t}{2(1+t^2)^{3/2}}\right\}\left(1+t^2\right)$$
$$=-\frac{t}{\sqrt{1+t^2}}=-\sin x$$

and similarly for $D_x\sin x$. The signs of cos x and sin x come from the definition and the signs of tan x: $\cos x>0$ $(-\frac{1}{2}\pi<x<\frac{1}{2}\pi)$, $\sin x<0$ $(-\frac{1}{2}\pi<x<0)$ and $\sin x>0$ $(0<x<\frac{1}{2}\pi)$. The signs of $D\cos x$ and $D\sin x$ follow and so the increasing/decreasing properties as stated. The addition formulae (*c*) are derived from (ii) for tan x:

$$\cos^2(x+y)=\frac{1}{1+\tan^2(x+y)}=\frac{(1-\tan x\tan y)^2}{(1-\tan x\tan y)^2+(\tan x+\tan y)^2}$$
$$=\frac{(1-\tan x\tan y)^2}{1+\tan^2x+\tan^2y+\tan^2x\tan^2y}=\frac{(1-\tan x\tan y)^2}{(1+\tan^2x)(1+\tan^2y)}$$

Put $\tan x=\dfrac{\sin x}{\cos x}$ and $\tan y=\dfrac{\sin y}{\cos y}$ and use

$$\sin^2x+\cos^2x=\sin^2y+\cos^2y=1:$$

$$\cos(x+y)=\sqrt{\frac{(\cos x\cos y-\sin x\sin y)^2}{(\sin^2x+\cos^2x)(\sin^2y+\cos^2y)}}=\cos x\cos y-\sin x\sin y.$$

The other addition formula follows similarly.

Both $\cos x$ and $\sin x$ can be defined on the closed interval $-\frac{1}{2}\pi \leqslant x \leqslant \frac{1}{2}\pi$. As $x \to \frac{1}{2}\pi$ and $t = \tan x \to \infty$, $\cos x = \dfrac{1}{\sqrt{1+t^2}} \to 0$ and $\sin x = \dfrac{t}{\sqrt{1+t^2}} \to 1$. Similarly, as $x \to -\frac{1}{2}\pi$, $\cos x \to 0$ and $\sin x \to -1$. So, write $\cos(\pm\frac{1}{2}\pi) = 0$ and $\sin(\frac{1}{2}\pi) = 1$, $\sin(-\frac{1}{2}\pi) = -1$. From the definition, $\cos 0 = 1$ and $\sin 0 = 0$. Hence the curves $y = \cos x$ and $y = \sin x$ are as shown in Fig. 15.8*b* for $-\frac{1}{2}\pi \leqslant x \leqslant \frac{1}{2}\pi$.

To extend the domain to all x is a matter of convention:

NOTATION: *cos* $(x+\pi) = -$*cos* x; *sin* $(x+\pi) = -$*sin* x;

$$tan\,(x+\pi) = tan\ x.$$

Hence, as shown in 12.5, the functions $y = \cos x$ and $y = \sin x$ are defined for all x and have period 2π. The function $y = \tan x$ is defined for all $x\left(x \neq \dfrac{2n+1}{2}\pi, \text{ for integral } n\right)$ and it has period π.

The expansion of the inverse tangent is obtained by integration term by term of the geometric series:

$$\frac{1}{1+x^2} = 1 - x^2 + x^4 - x^6 + \ldots + (-1)^n x^{2n} + \ldots$$

$$\tan^{-1} x = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \ldots + (-1)^n \frac{x^{2n+1}}{2n+1} + \ldots$$

absolutely and uniformly convergent for $|x| < 1$. The expansions of the cosine and sine functions are given by Taylor's Series:

$$\cos x = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} \ldots + (-1)^n \frac{x^{2n}}{(2n)!} + \ldots$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + \ldots$$

absolutely and uniformly convergent for all x.

15.9. Linear algebra. (Reference: Chapter 13.) The general concept of a vector space, 15.3(v) above, is developed:

(i) *Vector Space V over the Field F.* Write v for a vector of V and a for a scalar from F, denoting particular vectors and scalars by subscripts. The basic construction is that of linear dependence:

Definition: *The set* $\{v_1, v_2, \dots v_m\}$ *is* **linearly dependent** *if there is a set* $\{a_1, a_2, \dots a_m\}$ *of scalars, not all zero, so that* $\sum_{r=1}^{m} a_r v_r = 0$.

It follows that, if $\{v, v_1, v_2, \dots v_m\}$ are linearly dependent and such that $a \neq 0$ in the set of corresponding scalars $\{a, a_1, a_2, \dots a_m\}$, then v can be expressed as a *linear combination* of the set $\{v_1, v_2, \dots v_m\}$:

$$v = \sum_{r=1}^{m} \lambda_r v_r \quad \text{for some scalars } \lambda_1, \lambda_2, \dots \lambda_m.$$

It is only necessary to identify λ_r with a_r/a $(a \neq 0)$, and some of the λ's can be zero.* It also follows that a set $\{v_1, v_2, \dots v_m\}$ of m vectors is *linearly independent* if no set $\{a_1, a_2 \dots a_m\}$ of scalars exists for $\sum_{=1}^{m} a_r v_r \neq 0$, the case $a_1 = a_2 = \dots = a_m = 0$ being excluded. Hence, if the m vectors are linearly independent, then $\sum_{r=1}^{m} a_r v_r \neq 0$, for any scalars not all zero. No one vector of a linearly independent set is a linear combination of the others.

Definition: *The set* $S_m = \{v_1, v_2, \dots v_m\}$ **spans** *the vector space* V *if every vector of* V *is a linear combination of* S_m*:*

$$v = \sum_{r=1}^{m} \lambda_r v_r \quad \textit{for every } v \textit{ and for some scalars } \lambda_1, \lambda_2, \dots \lambda_m.$$

If there is no such set S_m, V is said to be of *infinite dimension*. Otherwise, from the several (at least one) integers m such that a set S_m spans V, pick out the smallest integer n. V is then said to have dimension n. If V is of dimension n, no set of fewer than n vectors spans V but there is at least one set of n vectors which does.

The main property of a vector space V of dimension n is:

Theorem: *If* $S_k = \{v_1, v_2, \dots v_k\}$ *is any set of vectors of* V *of dimension* n*:*

(i) $k < n$ *implies that* S_k *cannot span* V

(ii) $k > n$ *implies that* S_k *cannot be linearly independent*

(iii) $k = n$ *implies that* S_k *is linearly independent when it spans* V *and* S_k *spans* V *when it is linearly independent.*

* Indeed all the λ's can be zero, so that the zero vector $v = 0$ is a linear combination of any set of non-zero vectors, a rather trivial case.

Here (i) follows from the definition and (iii) is a consequence of the others; (ii) is the case which requires proof. We first prove the following proposition. Suppose that every vector of a set

$$S_k = \{v_1, v_2, \dots v_k\}$$

of k linearly independent vectors is a linear combination of a set $\{u_1, u_2, \dots u_n\}$ of n vectors:

$$v_i = \sum_{r=1}^{n} \lambda_{ir} u_r \quad \text{for some } \lambda_{ir} \quad (i=1, 2, \dots k;\ r=1, 2, \dots n) \quad \dots(1)$$

The proposition is that $k \leqslant n$. The proof is by mathematical induction on k. The proposition is obvious for $k=1$; if it holds for $k-1$, it must then be shown to hold for k. The induction hypothesis is: a set of $k-1$ linearly independent vectors is a linear combination of a set of n vectors and then $k-1 \leqslant n$, whatever n may be. There are two possibilities to consider. *First*, it may be that $\lambda_{1n} = \lambda_{2n} = \dots = \lambda_{kn} = 0$ so that u_n does not appear in (1). Hence each vector of S_k is a linear combination of $n-1$ vectors. Drop one of S_k to get a set of $k-1$ linearly independent vectors, each a linear combination of $n-1$ vectors. By the induction hypothesis, $k-1 \leqslant n-1$, i.e. $k \leqslant n$ and the induction is complete. *Second*, if at least one of these λ's is not zero, we can take it (by suitable ordering of vectors in S_k) as $\lambda_{kn} \neq 0$. Drop v_k from S_k to get the linearly independent set $S_{k-1} = \{v_1, v_2, \dots v_{k-1}\}$ of $k-1$ vectors. Write:

$$w_i = v_i - \frac{\lambda_{in}}{\lambda_{kn}} v_k \quad (i=1, 2, \dots k-1) \quad \dots\dots\dots\dots\dots\dots(2)$$

and assemble into a set $S_{k-1}' = \{w_1, w_2, \dots w_{k-1}\}$. Substitute (1) into (2) and notice that u_n disappears, leaving w_i as a linear combination of $\{u_1, u_2, \dots u_{n-1}\}$. Further, since the v's are linearly independent, (2) ensures that the w's are also. Hence the set S_{k-1}' has $k-1$ linearly independent vectors, each a linear combination of $n-1$ vectors. The induction hypothesis gives $k-1 \leqslant n-1$, i.e. $k \leqslant n$. Again the induction is complete. The proposition is established.

Proof of (ii): V is of dimension n and a set $\{u_1, u_2, \dots u_n\}$ of n vectors exists so that every v of V is a linear combination of them. If $S_k = \{v_1, v_2, \dots v_k\}$ is linearly independent, and (like all vectors) each is a linear combination of the n u's, then by our proposition $k \leqslant n$. Hence, if $k > n$, S_k is not linearly independent. Q.E.D.

From the theorem, we can say of a vector space V of dimension n that fewer than n vectors may be linearly independent but cannot span V; more than n vectors may span V but cannot be linearly independent. Both properties can hold only for a set of precisely n vectors. Hence:

DEFINITION: *A* **basis** *for V of dimension n is a set $S_n = \{v_1, v_2, \dots v_n\}$ of n vectors,* **both** *linearly independent* **and** *spanning V.*

A basis always exists; but it need not be unique. Usually there are many possible bases for V, all of precisely n vectors. To summarise:

> A vector space V over F of dimension n has at least one linearly independent set $S_n = \{v_1, v_2, \dots v_n\}$ as a basis. S_n spans V:
>
> $$v = \sum_{r=1}^{n} \lambda_r v_r \quad \text{for some scalars } \lambda_1, \lambda_2, \dots \lambda_n$$
>
> S_n is at the same time the *largest* of all linearly independent sets of V and the *smallest* of all sets spanning V.

Notice, in particular, that any set of $n+1$ vectors must be linearly dependent and one of them is a linear combination of the others; a set of $n-1$ vectors cannot span V and all vectors of V cannot be expressed as linear combinations of them.

(ii) *Space $V_n(F)$ of n-tuples.* Write $v = (x_1, x_2, \dots x_n)$ as an n-tuple of elements taken from the field F of scalars. Define vector addition: if $v = (x_1, x_2, \dots x_n)$ and $v' = (x_1', x_2', \dots x_n')$, then

$$v + v' = (x_1 + x_1', x_2 + x_2', \dots x_n + x_n').$$

Then the v's are an additive group, the zero vector being

$$0 = (0, 0, \dots 0).$$

Take a scalar a from F and define the scalar product:

$$\text{if } v = (x_1, x_2, \dots x_n), \text{ then } av = (ax_1, ax_2, \dots ax_n).$$

Then the v's form a vector space over F, denoted $V_n(F)$. A basis for $V_n(F)$, and hence its dimension, are easily found. Write the vectors:

$$\epsilon_1 = (1, 0, 0, \dots 0); \ \epsilon_2 = (0, 1, 0, \dots 0); \ \dots \ \epsilon_n = (0, 0, 0, \dots 1)$$

where ϵ_r has 1 in the rth place and 0's elsewhere. Then:

THEOREM: *The set $\{\epsilon_1, \epsilon_2, \dots \epsilon_n\}$ is a basis for the vector space $V_n(F)$ of n-tuples and $V_n(F)$ has dimension n.*

Proof: by the rules for sums and scalar products of n-tuples:

$$\begin{aligned}(x_1, x_2, \dots x_n) &= (x_1, 0, \dots 0) + (0, x_2, \dots 0) + \dots + (0, 0, \dots x_n)\\ &= x_1(1, 0, \dots 0) + x_2(0, 1, \dots 0) + \dots + x_n(0, 0, \dots 1)\\ &= x_1\epsilon_1 + x_2\epsilon_2 + \dots + x_n\epsilon_n\end{aligned}$$

i.e. the set $\epsilon_1, \epsilon_2, \dots \epsilon_n$ spans $V_n(F)$. Again:

$$\begin{aligned}\sum_{r=1}^{n} a_r\epsilon_r &= a_1(1, 0, \dots 0) + a_2(0, 1, \dots 0) + \dots + a_n(0, 0, \dots 1)\\ &= (a_1, 0, \dots 0) + (0, a_2, \dots 0) + \dots + (0, 0, \dots a_n)\\ &= (a_1, a_2, \dots a_n) \neq 0 \quad \text{(unless all } a_r = 0)\end{aligned}$$

i.e. the set $\epsilon_1, \epsilon_2, \dots \epsilon_n$ is linearly independent. It is a basis for $V_n(F)$ and, since it has n elements, $V_n(F)$ is of dimension n. Q.E.D.

This is not the only basis for $V_n(F)$. There are many others, e.g.

$$v_1 = (1, 0, 0, \dots 0);\ v_2 = (1, 1, 0, \dots 0);\ \dots\ v_n(1, 1, 1, \dots 1).$$

The importance of $V_n(F)$ stems from the result:

THEOREM: *Any vector space V over F of dimension n is isomorphic with $V_n(F)$, the isomorphism preserving sums and scalar products.*

Proof: a basis for V is some set $S_n = \{v_1, v_2, \dots v_n\}$ and any vector $v = \sum\limits_{r=1}^{n} \lambda_r v_r$ for some scalars $\lambda_1, \lambda_2, \dots \lambda_n$ from F. These scalars are unique. For, otherwise, let $v = \sum\limits_{r=1}^{n} \mu_r v_r$ also and $\sum\limits_{r=1}^{n} (\lambda_r - \mu_r)v_r = v - v = 0$ i.e. $\lambda_r = \mu_r$ all r. Write $\lambda = (\lambda_1, \lambda_2, \dots \lambda_n)$, a unique n-tuple of $V_n(F)$. Hence, to v of V, there corresponds a unique λ of $V_n(F)$. Conversely, given $\lambda = (\lambda_1, \lambda_2, \dots \lambda_n)$ of $V_n(F)$, then $\sum\limits_{r=1}^{n} \lambda_r v_r$ is a vector of V by the sum and scalar product rules. There is a one-one mapping, $v \leftrightarrow \lambda$, of V onto $V_n(F)$. The mapping preserves sums: if $v \leftrightarrow \lambda$ and $v' \leftrightarrow \lambda'$, then $v + v' = \sum\limits_{r=1}^{n} (\lambda_r + \lambda_r')v_r$ and $\lambda + \lambda' = (\lambda_1 + \lambda_1', \lambda_2 + \lambda_2', \dots \lambda_n + \lambda_n')$, i.e. $v + v' \leftrightarrow \lambda + \lambda'$. It preserves scalar products: if $v \leftrightarrow \lambda$ and a is a scalar, then $av = \sum\limits_{r=1}^{n} (a\lambda_r)v_r$ and $a\lambda = (a\lambda_1, a\lambda_1, \dots a\lambda_n)$, i.e. $av \leftrightarrow a\lambda$. The mapping is an isomorphism. Q.E.D.

In summarising, we allow first for the possibility that a vector space V is of infinite dimension. A case in point is the set $F[x]$ of polynomials over a field F, a ring (integral domain) under the operations of + and ×. Using the same field F for scalars, we define scalar products by (6) of 15.2 and we find that $F[x]$ is a set of vectors $v=(f_0, f_1, f_2, \ldots)$ over F. Here v is similar to the n-tuple $(x_1, x_2, \ldots x_n)$ except that there is no fixed n. Hence, the integral domain $F[x]$ is also a vector space over F of infinite dimension.

Otherwise, if V is of dimension n, the last theorem shows that V can be replaced for all algebraic purposes (up to isomorphism) by the space $V_n(F)$ of n-tuples. Hence, *any* vector space of finite dimension n can be effectively described solely by the integer n and the field F. For F provides *both* the scalars for scalar products *and* the constituents of the n-tuples of $V_n(F)$ equivalent to V. But $V_n(F)$ is still a wide concept. It can be specialised by taking F as the field of real numbers and by adding definitions of length and angle. $V_n(F)$ is then a space in the geometric sense. If length and angle are defined as in 8.4, $V_n(F)$ becomes the Euclidean space $E_n(F)$.

We have stressed that a ring (or field) and a vector space are two different systems of double composition. But, though different, they are not exclusive: a set S may be both a ring (or field) and a vector space. This is so if *three* operations, subject to the appropriate rules, are defined in S: sums, products and scalar products. An outside field F, different from S, is needed for scalars. Again, this does not rule out the possibility that F is a part of S. An example makes all this clear. The set C of complex numbers, $z=x+iy=$number pair (x, y), is a field under + and ×. Scalar multiplication of number pairs by real scalars a can be defined: $a(x, y)=(ax, ay)$, giving the complex number $(ax)+i(ay)$. Hence C is also a vector space, of dimension 2, over the field of real numbers.

(iii) *Linear Transformations.* A transformation, in general, is a mapping of one set into another. A *linear transformation* is the particular case of a mapping of one vector space V into another V', preserving the operations of addition and scalar multiplication. This is the same as saying that linear combinations of vectors are carried over from V to V' in the mapping. Hence the general and abstract concept:

DEFINITION: *If* $V=\{v_1, v_2, \ldots\}$ *and* $V'=\{v_1', v_2', \ldots\}$ *are two vector spaces over the same field* $F=\{\lambda_1, \lambda_2, \ldots\}$, *a* **linear transformation** *of* V

into V' is a mapping which carries a linear combination of vectors of V into the same linear combination of vectors of V':

$$\sum_{r=1}^{k} \lambda_r v_r \rightarrow \sum_{r=1}^{k} \lambda_r v_r' \quad \textit{if } v_r \rightarrow v_r' \ (r=1, 2, \ldots k).$$

The mapping may be one-one but generally it is many-one.

To bring the concept down to earth, take the case commonly used: two spaces $V_n(F)$ and $V_m(F)$ over the field F of real numbers. Write $x=(x_1, x_2, \ldots x_n)$ as an n-tuple of $V_n(F)$ and $y=(y_1, y_2, \ldots y_m)$ as an m-tuple of $V_m(F)$. The linear transformation is then: $x \xrightarrow[T]{} y$.

As a basis for $V_n(F)$ take $\{\epsilon_1, \epsilon_2, \ldots \epsilon_n\}$ where ϵ_s is the n-tuple with 1 in the sth place and 0's elsewhere. As a basis for $V_m(F)$ take $\{\eta_1, \eta_2, \ldots, \eta_m\}$ where η_r is the m-tuple with 1 in the rth place and 0's elsewhere. Then:

$$x = \sum_{s=1}^{n} x_s \epsilon_s \quad \text{and} \quad y = \sum_{r=1}^{m} y_r \eta_r \ldots\ldots\ldots\ldots\ldots\ldots\ldots(3)$$

Under T, each of $\epsilon_1, \epsilon_2, \ldots \epsilon_n$ as a vector of $V_n(F)$ has an image in $V_m(F)$, a certain m-tuple. With $s=1, 2, \ldots n$, write them:

$$\epsilon_s \rightarrow (a_{1s}, a_{2s}, \ldots a_{ms}) = a_{1s}\eta_1 + a_{2s}\eta_2 + \ldots + a_{ms}\eta_m.$$

There are $m \times n$ scalars a_{rs}, for $r=1, 2, \ldots m$ and $s=1, 2, \ldots n$. They are fixed by the specification of the linear transformation T. So:

$$\epsilon_s \rightarrow \sum_{r=1}^{m} a_{rs}\eta_r \quad s=1, 2, \ldots n \quad \ldots\ldots\ldots\ldots\ldots\ldots(4)$$

Let x and y correspond under T. By (3) and (4):

$$\sum_{s=1}^{n} x_s \epsilon_s \rightarrow \sum_{s=1}^{n} \sum_{r=1}^{m} x_s a_{rs} \eta_r = \sum_{r=1}^{m} \left(\sum_{s=1}^{n} a_{rs} x_s \right) \eta_r.$$

But
$$\sum_{s=1}^{n} x_s \epsilon_s \rightarrow \sum_{r=1}^{m} y_r \eta_r.$$

So:
$$y_r = \sum_{s=1}^{n} a_{rs} x_s \quad r=1, 2, \ldots m \quad \ldots\ldots\ldots\ldots\ldots\ldots(5)$$

The linear transformation T reduces to the m equations (5) and these determine $y_1, y_2, \ldots y_m$ when $x_1, x_2, \ldots x_n$ are given. Hence, the linear transformation is completely specified by the set of $m \times n$ scalars a_{rs}, i.e. by the matrix $\mathbf{A} = || \, a_{rs} \, ||$.

THEOREM: *A linear transformation T of $V_n(F)$ into $V_m(F)$ is described by a matrix* $\mathbf{A} = \| a_{rs} \|$ *of order $m \times n$ and T is:*

$$y_r = \sum_{s=1}^{n} a_{rs} x_s \quad r = 1, 2, \dots m$$

where T carries $(x_1, x_2, \dots x_n)$ of $V_n(F)$ into $(y_1, y_2, \dots y_m)$ of $V_m(F)$.

To determine **A** for a given transformation T, it is only necessary to find what m-tuples in $V_m(F)$ correspond to the basis $\epsilon_1, \epsilon_2, \dots \epsilon_n$ of $V_n(F)$. The constituents of the m-tuple corresponding to ϵ_1 are the first column of **A**, and so on. In the simplest case (as in 7.5) $n = m = 2$ and (5) are:

$$y_1 = a_{11}x_1 + a_{12}x_2 \quad \text{and} \quad y_2 = a_{21}x_1 + a_{22}x_2.$$

There is, however, no need that $n = m$ in a linear transformation.

(iv) *Matrices and Rank.* The set of all n-tuples from a field F is a vector space $V_n(F)$ of dimension n. A sub-set of n-tuples can still be a vector space in its own right, of dimension less than n. For example, the set of points in three dimensions forms a vector space $V_3(F)$ while the subset of points lying on a plane is a vector space $V_2(F)$. Apply this idea to rows and columns of a matrix $\mathbf{A} = \| a_{rs} \|$ of order $m \times n$ with elements from the field of real numbers.

The rows of **A** are n-tuples $v_r = (a_{r1}, a_{r2}, \dots a_{rn})$, $r = 1, 2, \dots m$. A vector space V is got by adding all linear combinations of v_r's. Let V be of dimension ρ. Then $\rho \leqslant n$ where $\rho = n$ only when V comprises all n-tuples of real numbers and where $\rho < n$ otherwise. Further, $\rho \leqslant m$ where $\rho = m$ only when the v_r's are linearly independent and $\rho < m$ otherwise. Hence $\rho \leqslant$ smaller of m, n. Here ρ is the *row rank* of the matrix **A**.

Alternatively, the columns of **A** are m-tuples $w_s = (a_{1s}, a_{2s}, \dots a_{ms})$, $s = 1, 2, \dots n$. Suppose the vector space W, got from all linear combinations of the w_s's, has dimension ω. Again, $\omega \leqslant$ smaller of m, n. Here ω is the *column rank* of **A**. The basic result is:

THEOREM: *A matrix* **A** *of order $m \times n$ has the same row and column rank ($\rho = \omega$), with ρ linearly independent and $m - \rho$ linearly dependent rows, with ρ linearly independent and $n - \rho$ linearly dependent columns.*

The *rank* of **A** is ρ and $\rho \leqslant m$, $\rho \leqslant n$. Proof:

Suppose $\rho < \omega$ and arrange **A** so that the first ρ rows and the first ω columns are linearly independent. From the n-tuple

$$v_r = (a_{r1}, a_{r2}, \dots a_{rn})$$

form the ω-tuple $\quad \bar{v}_r = (a_{r1}, a_{r2}, \dots a_{r\omega})$.

But: $$v_k = \sum_{r=1}^{\rho} \lambda_{rk} v_r \quad (k=1, 2, \dots m) \quad \dots\dots\dots\dots\dots(6)$$

for some λ's not all zero, expressing the linear dependence on the first ρ rows. Since (6) implies the same relations for the 'smaller' ω-tuples:

$$\bar{v}_k = \sum_{r=1}^{\rho} \lambda_{rk} \bar{v}_r \quad (k=1, 2, \dots m) \dots\dots\dots\dots\dots(7)$$

Write: $$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \dots + a_{1\omega}x_\omega = 0 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2\omega}x_\omega = 0 \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ a_{\rho 1}x_1 + a_{\rho 2}x_2 + \dots + a_{\rho\omega}x_\omega = 0 \end{array}\right\} \dots\dots\dots\dots\dots(8)$$

as ρ equations in ω variables $x_1, x_2, \dots x_\varpi$. Given $\rho < \omega$ and from 13.8 above, x's not all zero exist as a solution of (8), a non-zero value of (at least) one of the x's being assigned in advance. From (7) and (8):

$$a_{k1}x_1 + a_{k2}x_2 + \dots + a_{k\omega}x_\omega = \left(\sum_{r=1}^{\rho} \lambda_{rk} a_{r1}\right)x_1 + \left(\sum_{r=1}^{\rho} \lambda_{rk} a_{r2}\right)x_2 + \dots$$
$$+ \left(\sum_{r=1}^{\rho} \lambda_{rk} a_{r\omega}\right)x_\omega$$
$$= \sum_{r=1}^{\rho} \lambda_{rk}(a_{r1}x_1 + a_{r2}x_2 + \dots + a_{r\omega}x_\omega) = 0$$

for some x's not all zero and for any $k = 1, 2, \dots m$. Hence (8) extends to a full set of m equations in the x's. Consequently, the first ω columns of **A** are linearly dependent, the multiples being the x's. But ω is defined so that the first ω columns of **A** are linearly independent. This is a contradiction. Hence $\rho \nless \omega$. Similarly, $\omega \nless \rho$. Hence $\rho = \omega$. The rest of the theorem follows at once. Q.E.D.

APPENDIX

FORMULAE OF ELEMENTARY ALGEBRA AND TRIGONOMETRY

No detailed knowledge of elementary mathematics is assumed in this text but certain simple results are needed from time to time. The formulae are developed in this Appendix. Where proofs are not given, they are to be found in the standard school texts.

A.1. Powers and exponents. Let a be a positive real number and n a positive integer. Then a^n is a notation for $a \times a \times \ldots \times a$ (n times). From the definition, the following properties follow immediately:

$$a^m a^n = a^{m+n};\ (a^m)^n = a^{mn};\ (ab)^n = a^n b^n$$

for any positive real numbers a and b and any positive integers m and n. An extension of the notation permits the expression a^x to be written for any positive real number a and any rational x. For positive fractional x, a^x is defined as a root of a; when x is negative, a^x is defined as a reciprocal. As special cases, a^1 is taken as a and a^0 stands for unity*. The complete notation can be spelled out:

NOTATION: *The expression a^x is defined for a positive real number a and various rational values of x:*

$x = n$, *where n is a positive integer:*

$$a^n = a \times a \times \ldots \times a \quad (n \textit{ times})$$

$x = 0$: $a^0 = 1$

* It may be said that a^0 *must* be 1 since $a^m a^n = a^{m+n}$ with $m = 0$ gives $a^0 a^n = a^{0+n} = a^n$ so that $a^0 = 1$. Similarly it may be said (e.g.) that $a^{\frac{1}{2}}$ *must* be $\sqrt{a}$, since $(a^m)^n = a^{mn}$ with $m = \frac{1}{2}$ and $n = 2$ gives $(a^{\frac{1}{2}})^2 = a^{\frac{1}{2}2} = a$ so that $a^{\frac{1}{2}} = \sqrt{a}$. This is a misconception. We are at liberty to define a^0 and $a^{\frac{1}{2}}$ in any way we like; there is no 'must' about it. What we *choose* to do is to take $a^0 = 1$ to preserve the rule $a^m a^n = a^{m+n}$ and $a^{\frac{1}{2}} = \sqrt{a}$ to preserve the rule $(a^m)^n = a^{mn}$. We could choose otherwise, e.g. $a^0 = 0$ and $a^{\frac{1}{2}} = 1/a^2$, but it would be wasteful. Generalisation in mathematics aims at preserving, rather than scrapping, existing rules.

$x=p/q$, *where p and q are positive integers:*

$$a^{p/q}=\sqrt[q]{a^p}=\textit{positive qth root of } a^p$$

$x=-r$, *where r is a positive rational:*

$$a^{-r}=\frac{1}{a^r}=\textit{reciprocal of } a^r.$$

Here a^x is the xth *power* of a and x is the *exponent* of the power.

The limitation $a>0$ can be relaxed for some but not for all exponents. It is in order to write a^x for $a<0$ when x is integral, but not always when x is fractional. Powers like $a^3=a\times a\times a$ and $a^{-2}=\dfrac{1}{a\times a}$ are valid for negative a. On the other hand $a^{1/2}=\sqrt{a}$ has no meaning (within the domain of real numbers) if a is negative, though $a^{1/3}=\sqrt[3]{a}$ still holds. It is correct, moreover, to write $a^x=0$ when $a=0$, except when x is a negative rational. For example, $0^{1/2}=\sqrt{0}=0$ but $0^{-1/2}=\dfrac{1}{\sqrt{0}}$ has no meaning.

The properties given above for positive integral exponents remain true for any rational exponents. The notation is designed to achieve this:

$$a^xa^y=a^{x+y};\ (a^x)^y=a^{xy};\ (ab)^x=a^xb^x \quad \ldots\ldots\ldots\ldots\ldots(1)$$

for any positive real numbers a and b and for any rational x and y. Various applications of the properties (1) are involved in the examples:

$$\sqrt{a^3}\sqrt{a^5}=a^{3/2}a^{5/2}=a^{3/2+5/2}=a^4$$

$$\sqrt{a^3}\sqrt{a^5}=\sqrt{a^3a^5}=\sqrt{a^{3+5}}=\sqrt{a^8}=a^4$$

$$\frac{(ab)^3}{a^5}=\frac{a^3b^3}{a^5}=b^3(a^3a^{-5})=b^3a^{3-5}=b^3a^{-2}=\frac{b^3}{a^2}$$

$$\frac{\sqrt[3]{a}}{\sqrt{a}}=a^{1/3}a^{-1/2}=a^{-1/6}=\frac{1}{\sqrt[6]{a}}.$$

A.2. Logarithms. A logarithm is the inverse of a power. If a is a positive real number, then so is a^y for any rational y (positive, negative or zero).

NOTATION: *If $x=a^y$, then $y=\log_a x$, read 'logarithm of x to base a', defined for suitable positive real values of x.*

Since a logarithm is, by definition, an exponent, the properties (1) of powers can be used to derive properties of logarithms:

$$\log_a(xy) = \log_a x + \log_a y;\ \log_a(x^b) = b \log_a x;\ \log_a x = \log_a b \log_b x \quad (2)$$

The last of (2) implies that the logarithms of various x to one base a are simply re-scalings of the corresponding logarithms to a second base b. Given a set of logarithms to the base b, the corresponding set of logarithms to the base a are to be written by multiplying through by the constant factor $\log_a b$. It is for this reason that the particular base chosen for logarithms is of little importance; any convenient base will do. *Common logarithms* as used in arithmetical work are logarithms to the base 10, a very convenient base. So $y = \log_{10} x$ implies simply: $x = 10^y$.

There is a difficulty here: what do we mean by 'suitable' x in the definition of the logarithm of x? Since $y = \log_a x$ means $x = a^y$ and since exponents are (so far) limited to rationals, it follows that only rational logarithms can be written as yet. So $\log_a x = y$ must be rational and x must be a real number which is a rational power: $x = a^y$. This is a very severe limitation.

Consider common logarithms and drop for convenience the reference to the base 10. Then $\log x = y$ means $x = 10^y$ only for rational y. Hence x must be of the form: $x = 10^{p/q} = \sqrt[q]{10^p}$ (p and q integers, $q > 0$). So $0{\cdot}01 = 10^{-2}$ has logarithm $\log 0{\cdot}01 = -2$, and $\sqrt{10} = 10^{0\cdot5}$ has logarithm $\log \sqrt{10} = 0{\cdot}5$. But, in fact, most real numbers x have no (rational) logarithm. For example, take the positive integers greater than 1. The only integers among 2, 3, 4, 5, ... with (rational) logarithms are multiples of 10. The integers 10, 100, 1000, ... have logarithms 1, 2, 3, ...; any other integer k has no (rational) logarithm. To see this: suppose $\log k = p/q$ (rational, p and q positive integers) so that $k = 10^{p/q} = \sqrt[q]{10^p}$ and $k^q = 10^p$ = number with 0 as last digit in the decimal system. This is impossible; no integral power of k can have 0 in the last digit. Powers of 2 end in 2, 4, 6 or 8, powers of 3 in 1, 3, 7 or 9, and so on. Hence $\log k$ cannot be rational.

There is no (rational) $\log x$ for values of x as simple as 2 or 3. Nor are we able, as yet, to wriggle out by saying that log 2 or log 3 is irrational; we have no definition of irrational powers of 10. We are in an awkward situation; we cannot even justify the use of tables of logarithms in ordinary arithmetic. The tables show for example:

log 2 = 0·30103 .... This is not a rational number (no rational power of 10 gives 2). We are not entitled to take it as an irrational number (no irrational powers of 10 are defined). The missing link is the definition of irrational powers and it can be supplied — in Chapter 12 after much water has passed under the bridges. It turns out to be true that log 2 is irrational, approximated by the rational 0·30103 to five decimal places. But this involves a surprisingly sophisticated concept.

A.3. Roots of polynomial equations. We are given a polynomial equation with real coefficients and of degree n, where n is a positive integer. We wish to solve it in the sense of finding values for all the roots. Moreover, we require a solution which is a general process or formula, applicable to all equations of given degree n, a process which involves only the operations of addition, subtraction, multiplication and division, together with root extraction (square roots, cube roots, and so on). The position finally established by the work of Galois (1811–32) is a very remarkable one. There are general processes for $n = 2$, 3 and 4, a simple formula for the quadratic, much more difficult processes for cubics and quartics — but it is just not possible to solve all polynomial equations of degree $n > 5$.

The general process for a quadratic is that of 'completing the square':

Given: $$ax^2 + bx + c = 0 \quad (a \neq 0)$$

write: $$a\left\{x^2 + 2\frac{b}{2a}x + \left(\frac{b}{2a}\right)^2\right\} = \frac{b^2}{4a} - c$$

i.e. $$\left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2}.$$

Make the transformation to another variable y: $x = y - \frac{b}{2a}$. Then:

$$y^2 = \frac{b^2 - 4ac}{4a^2} \quad \text{i.e.} \quad y = \pm\frac{1}{2a}\sqrt{b^2 - 4ac}.$$

So: $$x = \frac{1}{2a}\left(-b \pm \sqrt{b^2 - 4ac}\right) \quad \text{.......................(3)}$$

The general formula (3) applies to any quadratic $ax^2 + bx + c = 0$ with real coefficients a, b and c. The only requirement $(a \neq 0)$ is needed to ensure that the polynomial is indeed a quadratic. If $b^2 > 4ac$, the two roots (3) are real and distinct (rational if b^2-4ac is a perfect square); if $b^2=4ac$, there is a double root which must be rational $\left(-\frac{b}{2a}\right)$. On the other hand, if $b^2<4ac$, then the two roots (3) are conjugate complex.

To illustrate that a general but complicated process can be obtained for the solution of a cubic or quartic equation, consider the following development for the cubic. It involves more than one transformation: The first step is that of 'completing the cube', to get rid of x^2:

Given: $$ax^3+bx^2+cx+d=0 \quad (a\neq 0)$$

divide through by a and arrange:

$$x^3+3\frac{b}{3a}x^2+3\left(\frac{b}{3a}\right)^2x+\left(\frac{b}{3a}\right)^3=\left(\frac{b^2}{3a^2}-\frac{c}{a}\right)x+\left(\frac{b^3}{27a^3}-\frac{d}{a}\right)$$

i.e. $$\left(x+\frac{b}{3a}\right)^3=\left(\frac{b^2}{3a^2}-\frac{c}{a}\right)\left(x+\frac{b}{3a}\right)-\frac{b}{3a}\left(\frac{b^2}{3a^2}-\frac{c}{a}\right)+\left(\frac{b^3}{27a^3}-\frac{d}{a}\right).$$

Make the transformation $x=y-\frac{b}{3a}$ so that the equation becomes:

$$y^3=\alpha y+\beta$$

where $$\alpha=\frac{b^2}{3a^2}-\frac{c}{a} \quad \text{and} \quad \beta=\frac{bc}{3a^2}-\frac{2b^3}{27a^3}-\frac{d}{a}.$$

The next step is to get rid of the term in y. Make the transformation $y=z+\frac{\alpha}{3z}$ $(z\neq 0)$ as originally devised by Vieta (1540–1603):

$$\left(z+\frac{\alpha}{3z}\right)^3=\alpha\left(z+\frac{\alpha}{3z}\right)+\beta$$

giving: $$z^3-\beta+\frac{\alpha^3}{27z^3}=0.$$

Finally, multiply through by z^3 and transform by $w=z^3$:

$$w^2-\beta w+\frac{\alpha^3}{27}=0.$$

This can be solved: $$w=\tfrac{1}{2}\left(\beta\pm\sqrt{\beta^2-\frac{4\alpha^3}{27}}\right).$$

Jobbing backwards, we obtain z, y and then x in succession. For each of the two values of w, we write the three cube roots (as in 3.8) to get

z. For each of these, we get $y=z+\frac{\alpha}{3z}$ and finally $x=y-\frac{b}{3a}$. The result is $2\times 3=6$ values of x but these are found to be equal in pairs. The outcome is a general process for obtaining the three roots of the original cubic; two of the roots may be conjugate complex. The tricky step is getting the three cube roots of w to provide z.

A.4. Solution of two linear equations. As an extension, we can attempt to solve two polynomial equations, each involving two variables. The simplest case is that of linear equations:

$$a_1x+b_1y+c_1=0 \quad \text{and} \quad a_2x+b_2y+c_2=0$$

where all six coefficients have real values and where the variables x and y are real. Suppose that $a_1b_2-a_2b_1\neq 0$. If $b_2\neq 0$, eliminate y by use of the second equation:

$$a_1x+b_1\left(-\frac{a_2x+c_2}{b_2}\right)+c_1=0$$

i.e. $$x=\frac{b_1c_2-b_2c_1}{a_1b_2-a_2b_1} \quad \text{and} \quad y=-\frac{a_2}{b_2}x-\frac{c_2}{b_2}=\frac{c_1a_2-c_2a_1}{a_1b_2-a_2b_1}\,.$$

So: $$\frac{x}{b_1c_2-b_2c_1}=\frac{y}{c_1a_2-c_2a_1}=\frac{1}{a_1b_2-a_2b_1} \quad (a_1b_2-a_2b_1\neq 0) \;\ldots\ldots(4)$$

On the other hand, if $b_2=0$, then $b_1\neq 0$ to ensure that $a_1b_2-a_2b_1\neq 0$. In this case, y can be eliminated by use of the first equation and the same result (4) follows. Hence, the equations have a unique solution (4), provided that $a_1b_2-a_2b_1\neq 0$.

This is the *main case*. The difficulty is that there are other cases, which may be called *degenerate cases*, arising when $a_1b_2-a_2b_1=0$. Write (4) as $\frac{x}{A}=\frac{y}{B}=\frac{1}{C}$, where $A=b_1c_2-b_2c_1$, $B=c_1a_2-c_2a_1$ and $C=a_1b_2-a_2b_1$. The main case has $C\neq 0$ and it does not matter whether either (or both) of A and B is zero. If $A=0$, then $x=\frac{A}{C}=0$; if $B=0$, then $y=\frac{B}{C}=0$. There are two degenerate cases, with $C=0$.

(i) $C=0$ and either $A\neq 0$ or $B\neq 0$ (or both).

Speaking roughly, we say that (4) gives $x=\frac{A}{0}$ or $y=\frac{B}{0}$ (or both), i.e.

that x or y is 'infinite'. More strictly: $a_2=\lambda a_1$, $b_2=\lambda b_1$ but $c_2\neq\lambda c_1$ for some multiple λ. The equations are inconsistent, shown graphically by parallel lines.

(ii) $A=B=C=0$.

Again, speaking roughly, we say that (4) gives each of x and y in the form $\frac{0}{0}$ and that x and y are 'indeterminate'. More strictly, $a_2=\lambda a_1$, $b_2=\lambda b_1$ and $c_2=\lambda c_1$ for some multiple λ. The equations are identical, shown graphically by coincident lines.

A.5. Completing the square. An essential property of the system of real numbers is that an expression in real variables which is a perfect square must be non-negative. It is zero if the terms inside the square vanish; otherwise it is positive. Further, if an expression is the sum of two or more perfect squares, then it must be positive, except that it is zero when all the squares separately vanish. This property is not necessarily true of other number systems; with complex numbers, for example, a perfect square can be negative: $i^2=-1$.

Much use of this property of real numbers is made in algebra. For example, completing the square helps in a quadratic polynomial:

$$2x^2-x-3=2(x^2-\tfrac{1}{2}x+\tfrac{1}{16})-3-\tfrac{1}{8}=2(x-\tfrac{1}{4})^2-\tfrac{25}{8}.$$

Hence: $$(2x^2-x-3)-(-\tfrac{25}{8})=2(x-\tfrac{1}{4})^2\geqslant 0$$

i.e. $2x^2-x-3$ is greater than $(-\frac{25}{8})$, except that it equals $-\frac{25}{8}$ when $x=\frac{1}{4}$. The quadratic expression has the smallest value $(-\frac{25}{8})$ when $x=\frac{1}{4}$. To generalise:

Given: $$y=ax^2+bx+c \quad (a>0)$$

write $$y=a\left(x^2+\frac{b}{a}x+\frac{b^2}{4a^2}\right)+c-\frac{b^2}{4a}=a\left(x+\frac{b}{2a}\right)^2-\frac{b^2-4ac}{4a}$$

i.e. $$y-\left(-\frac{b^2-4ac}{4a}\right)=a\left(x+\frac{b}{2a}\right)^2\geqslant 0$$

i.e. $$y \text{ has minimum } -\frac{b^2-4ac}{4a} \text{ when } x=-\frac{b}{2a}.$$

As another example of reducing a difference to a square, we show that $\dfrac{x+y}{2}\geqslant\sqrt{(xy)}$ for all positive real x and y (equals only if $x=y$) by writing the difference:

$$\frac{x+y}{2}-\sqrt{(xy)}=\tfrac{1}{2}\{(\sqrt{x})^2+(\sqrt{y})^2-2\sqrt{x}\sqrt{y}\}=\tfrac{1}{2}(\sqrt{x}-\sqrt{y})^2\geqslant 0.$$

Squaring up is a useful process. It is, however, not reversible; the square of a is a^2 but the square root of a^2 is either a or $-a$. To illustrate this disadvantage, we can attempt to solve $2x-1+\surd(2x+1)=0$:

$$2x-1=-\surd(2x+1).$$

Square: $\quad 4x^2-4x+1=2x+1 \quad$ i.e. $\quad 2x(2x-3)=0.$

Hence $x=0$ and $x=3/2$ are possible. Checking, we find that $x=0$ does satisfy the original equation, but not $x=3/2$. However,

$$2x-1-\surd(2x+1)=0,$$

a different equation, squares up to the same quadratic; $x=3/2$ satisfies the equation, but not $x=0$.

A.6. Clearing the denominator. The numerator in a ratio is more easily handled than the denominator. The device of clearing the denominator of unwanted elements (e.g. roots) is a very useful one in algebra. Simple examples illustrate:

$$\frac{1}{\surd 2+1}=\frac{\surd 2-1}{(\surd 2+1)(\surd 2-1)}=\frac{\surd 2-1}{(\surd 2)^2-1}=\frac{\surd 2-1}{2-1}=\surd 2-1$$

$$\frac{\surd 2+1}{\surd 2-1}=\frac{(\surd 2+1)^2}{(\surd 2-1)(\surd 2+1)}=\frac{(\surd 2)^2+2\surd 2+1}{(\surd 2)^2-1}=2\surd 2+3$$

both depending on the simple result: $(\surd 2+1)(\surd 2-1)=(\surd 2)^2-1=1$. The trick is to multiply numerator and denominator by the same expression, so chosen to clear the denominator of the awkward $\surd 2$. More generally:

$$\frac{a+b\surd 2}{c+d\surd 2}=\frac{(a+b\surd 2)(c-d\surd 2)}{(c+d\surd 2)(c-d\surd 2)}=\frac{ac+bc\surd 2-ad\surd 2-2bd}{c^2-2d^2}$$

i.e. $$\frac{a+b\surd 2}{c+d\surd 2}=\left(\frac{ac-2bd}{c^2-2d^2}\right)+\left(\frac{bc-ad}{c^2-2d^2}\right)\surd 2 \qquad \text{.........................(5)}$$

The result (5) shows that the ratio of two numbers of the form $a+b\surd 2$, where a and b are rationals, is a number of the same form.

The same device works in many other cases. For example:

$$\frac{1}{\sqrt{x+1}+\surd x}=\frac{\sqrt{x+1}-\surd x}{(\sqrt{x+1})^2-(\surd x)^2}=\sqrt{x+1}-\surd x$$

if x is real and positive. Since $\sqrt{x+1}>\surd x$, so $\sqrt{x+1}+\surd x>2\surd x$ and:

$$\sqrt{x+1}-\surd x=\frac{1}{\sqrt{x+1}+\surd x}<\frac{1}{2\surd x} \quad \text{(all positive real } x\text{).}$$

The device is of particular use in handling ratios of complex numbers:

$$\frac{a+ib}{c+id}=\frac{(a+ib)(c-id)}{(c+id)(c-id)}=\frac{ac+ibc-iad-i^2bd}{c^2-(id)^2} \quad (i^2=-1)$$

i.e. $$\frac{a+ib}{c+id}=\left(\frac{ac+bd}{c^2+d^2}\right)+i\left(\frac{bc-ad}{c^2+d^2}\right) \quad \text{..................(6)}$$

The similarity between (5) and (6) is clear. In (6), the denominator is cleared of the awkward i. It shows that the ratio of two complex numbers is also a complex number.

A.7. Trigonometric Ratios

(i) In elementary geometry and trigonometry, an angle θ is measured in degrees. Consider first an *acute angle* θ, i.e. θ positive and less than a right angle ($0° < \theta < 90°$). The definition of the trigonometric ratios is given in terms of the triangle OPM of Fig. A.7, any right-angled triangle with $\angle POM = \theta$:

DEFINITION: *The trigonometric ratios (cosine, sine and tangent) of the acute angle θ are:*

$$\cos\theta=\frac{OM}{OP},\ \sin\theta=\frac{MP}{OP},\ \tan\theta=\frac{MP}{OM}.$$

Any right-angled triangle of the shape of OPM will do since all such triangles are similar, i.e. the ratios of sides are the same for all.

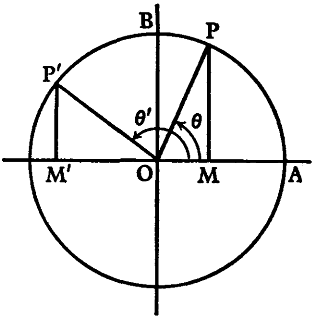


FIG. A. 7

It follows immediately that tan θ is the ratio of sin θ to cos θ and (from Pythagoras' Theorem $OP^2 = OM^2 + MP^2$) that the sum of the squares of sin θ and cos θ is 1. As a convenient notation, write $(\sin\theta)^2 = \sin^2\theta$ and similarly for the others. Hence:

$$\tan\theta=\frac{\sin\theta}{\cos\theta} \quad \text{and} \quad \sin^2\theta+\cos^2\theta=1 \quad \text{..............(7)}$$

Three trigonometric ratios are defined here and it is useful to have all three. But the results (7) show that any two can be expressed in terms of the other. Perhaps the simplest expression of this is:

$$\cos\theta=\frac{1}{\sqrt{1+t^2}} \quad \text{and} \quad \sin\theta=\frac{t}{\sqrt{1+t^2}} \quad \text{where } t=\tan\theta.$$

Values of the ratios for particular angles are obtained by drawing appropriate triangles. Limits as $\theta \to 0°$ or $\theta \to 90°$ can be written. So:

$\theta =$	0°	30°	45°	60°	90°
$\cos\theta$	1	$\frac{\sqrt{3}}{2}$	$\frac{1}{\sqrt{2}}$	$\frac{1}{2}$	0
$\sin\theta$	0	$\frac{1}{2}$	$\frac{1}{\sqrt{2}}$	$\frac{\sqrt{3}}{2}$	1
$\tan\theta$	0	$\frac{1}{\sqrt{3}}$	1	$\sqrt{3}$	*

* $\tan\theta \to \infty$ as $\theta \to 90°$.

The addition formulae are:

$$\left.\begin{aligned}
&\cos(\theta_1+\theta_2) &&\cos(\theta_1-\theta_2)\\
&\quad=\cos\theta_1\cos\theta_2-\sin\theta_1\sin\theta_2 &&\quad=\cos\theta_1\cos\theta_2+\sin\theta_1\sin\theta_2\\
&\sin(\theta_1+\theta_2) &&\sin(\theta_1-\theta_2)\\
&\quad=\sin\theta_1\cos\theta_2+\cos\theta_1\sin\theta_2 &&\quad=\sin\theta_1\cos\theta_2-\cos\theta_1\sin\theta_2\\
&\tan(\theta_1+\theta_2) &&\tan(\theta_1-\theta_2)\\
&\quad=\frac{\tan\theta_1+\tan\theta_2}{1-\tan\theta_1\tan\theta_2} &&\quad=\frac{\tan\theta_1-\tan\theta_2}{1+\tan\theta_1\tan\theta_2}
\end{aligned}\right\}$$

.........(8)

(ii) For a *positive angle* θ of any size, measured in degrees, the definition of the trigonometric ratios needs extension. In Fig. A.7, let P rotate anticlockwise around a circle centred at O describing an angle θ from the starting position A. Take signed distances from O along OA, positive to the right and negative to the left of O; take signed distances vertically (parallel to OB), positive upwards and negative downwards. Then proceed by stages:

When P is in the second quadrant (as P' in Fig. A.7), θ varies between 90° and 180°. Keep the definition of the trigonometric ratios, except that OP is positive but other lengths have a sign (+ or −). So, for θ' of Fig. A.7, we have

$$\cos\theta' = \frac{OM'}{OP'} = \text{negative} \quad \sin\theta' = \frac{M'P'}{OP'} = \text{positive}$$

and

$$\tan\theta' = \frac{\sin\theta'}{\cos\theta'} = \frac{M'P'}{OM'} = \text{negative.}$$

This is equivalent (as can be seen from Fig. A.7 with $\theta' = 180 - \theta$) to writing the following for θ acute, $180° - \theta$ in the second quadrant:

$$\cos(180° - \theta) = -\cos\theta \quad \text{and} \quad \sin(180° - \theta) = \sin\theta.$$

The extension to the third quadrant follows, equivalent to:

$$\cos(180° + \theta) = -\cos\theta \quad \text{and} \quad \sin(180° + \theta) = -\sin\theta$$

and so to the fourth quadrant:

$$\cos(360° - \theta) = \cos\theta \quad \text{and} \quad \sin(360° - \theta) = -\sin\theta.$$

This takes care of positive angles up to 360°. For still larger angles, P has rotated around the complete circle one or more times. Then:

$$\cos(n360° + \theta) = \cos\theta \quad \text{and} \quad \sin(n360° + \theta) = \sin\theta$$

for any positive integer n.

Finally, for a *negative angle* θ, the rotation of P is taken from A in the negative or clockwise direction. Then:

$$\cos(-\theta) = \cos\theta \quad \text{and} \quad \sin(-\theta) = -\sin\theta$$

sometimes expressed by saying that $\cos\theta$ is an 'even' function and $\sin\theta$ is an 'odd' function.

The results (7) and (8) hold for angles of any size. Indeed, the addition formulae for $\theta_1 - \theta_2$ in (8) follow from those for $\theta_1 + \theta_2$ by substituting $(-\theta_2)$ for θ_2.

A.8. Triangles. Denote the angles of a triangle by A, B and C and the opposite sides by a, b and c respectively. The relations between sides and angles are:

$$\frac{a}{\sin A} = \frac{b}{\sin B} = \frac{c}{\sin C} \quad \text{and} \quad a^2 = b^2 + c^2 - 2bc\cos A \ldots\ldots\ldots(9)$$

The second relation of (9) has two similar forms, for b^2 and for c^2 respectively. As a check, put $A = 90°$ so that $\sin A = 1$ and $\cos A = 0$. The first relations of (9) give $\sin B = \frac{b}{a}$ and $\sin C = \frac{c}{a}$, the definition of these trigonometric ratios. From the second relation of (9):

$$a^2 = b^2 + c^2,$$

and this is Pythagoras' Theorem for a right-angled triangle.

Two triangles are *similar* if corresponding angles are equal: $A = A'$, $B = B'$ and $C = C'$. From the first relations of (9), sides are proportional, i.e. $a : b : c = a' : b' : c'$. As a particular case, if corresponding sides are also equal, then the triangles are *congruent*.

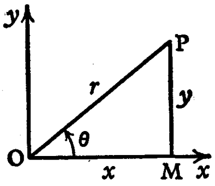
Necessary and sufficient conditions for congruent triangles are: (i) three sides correspond, or (ii) two sides and included angle correspond, or (iii) one side and two angles correspond.

The area $\triangle$ of a triangle ABC, defined as half the product of the base and the height, can be expressed in terms of sides and angles:

$$\triangle = \tfrac{1}{2}bc \sin A = \tfrac{1}{2}ca \sin B = \tfrac{1}{2}ab \sin C \quad \text{...........(10)}$$

The equality of the three expressions for $\triangle$ in (10) is ensured by the first relations of (9). The common value of the ratios $\dfrac{a}{\sin A}$, $\dfrac{b}{\sin B}$ and $\dfrac{c}{\sin C}$ is then seen to be $\dfrac{abc}{2\triangle}$.

A.9. Cartesian and polar co-ordinates. Fix axes Ox and Oy in a plane, each being a directed line on which a suitable scale of measurement is taken. In Fig. A.9, Ox is drawn horizontally to the right and Oy vertically upwards. Any point P in the plane needs two *co-ordinates* to fix it, corresponding to the two *dimensions* of the plane. Alternative systems of pairs of co-ordinates can be devised and two of them are in common use. In one system, that of *Cartesian co-ordinates*, the number pair (x, y) is attached to a point P, where x is the distance OM and y is the distance MP in Fig. A.9. Here each of x and y is any real number (positive, negative or zero). So, if $x>0$, P is to the right of Oy; if $x=0$, P is on Oy; if $x<0$, P is to the left of Oy. Similarly, the sign of y determines whether P is above, on or below Ox.



F^oG. A. 9

The other system, that of *polar co-ordinates*, attaches the number pair (r, θ) to P, where r is the distance OP and θ is the angle (in degrees) which OP makes anti-clockwise with Ox. Both co-ordinates are real numbers, r being zero or positive and θ lying in the interval $0° \leqslant \theta < 360°$.

The relations between the co-ordinate systems follow at once:

$$x = r \cos \theta \quad \text{and} \quad y = r \sin \theta \quad \text{.................(11)}$$

This is a transformation from given polar co-ordinates (r, θ) to corresponding Cartesian co-ordinates. The inverse transformation, also needed, is more difficult to specify. By use of (7):

$$r^2 = x^2 + y^2 \quad \text{and} \quad \tan\theta = \frac{y}{x}$$

which are also checked from Fig. A.9. Hence:

$$r = \sqrt{x^2 + y^2} \quad \text{and} \quad \theta = \tan^{-1}\frac{y}{x} \quad \ldots\ldots\ldots\ldots\ldots\ldots(12)$$

Since r is positive, the value of r in (12) is unambiguous, the positive square root of the positive expression $x^2 + y^2$ (apart from the special case $x = y = 0$ with $r = 0$). The value of θ is not unambiguous as shown in (12) since there are two angles in the interval $0° \leqslant \theta < 360°$ with a given value for $\tan\theta$. (Fig. 12.7*b* shows how $\tan\theta$ repeats itself in this interval.) One of these two values is to be selected and the other rejected. The criterion is: find r and select θ so that $\cos\theta = x/r$ (with sign of x) *and* so that $\sin\theta = y/r$ (with sign of y).

Hence (12) must be read subject to the condition:

> Of the two values of θ in the range $0° \leqslant \theta < 360°$ with $\tan\theta = y/x$, that value is taken for which $\cos\theta$ has the sign of x and $\sin\theta$ the sign of y.

The other value of θ in the range is to be rejected; it is such that $\cos\theta$ has the opposite sign to x and $\sin\theta$ the opposite sign to y.

EXERCISES: SOLUTIONS

1.9 **1.** $y = 32 + 1{\cdot}8x$ for $0 \leqslant x \leqslant 100$.

2.

$x =$	1	2	3	4
$y =$	4	6	4	1

3. $x = 2$. **4.** $x = \surd 3$; $\surd 3 < x \leqslant 2$.

10. Roots: 1 (twice) and $-\frac{1}{2}$; 0 (twice) and $\frac{3}{2}$.

13. $x = \frac{1}{2}(1-k)$, $y = \frac{1}{2}(1+k)$.

2.9 **1.** No.

14. Use Ex. 13 with $\theta_1 = \theta_2 = \theta$.

17. $x = 1$.

20.

$+$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

$\times$	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

3.9 **4.** Note that the second gives $\surd i = \pm(1+i)/\surd 2$ (2.9 Ex. 15).

8. All real $x (\neq \pm 1)$ for each rational fraction; but all real $x(\neq -1)$ for second reduced form.

11. Of quadratics $x^2 + ax + b$ where $(a, b) = (0, 0), (0, 1), (1, 0)$ or $(1, 1)$, only $x^2 + x + 1$ has no factors.

13. If α real (complex) then $g(x)$ has real (complex) coefficients. Other roots: -1; none; 1; none.

15. ± 1 and $(1 \pm i\surd 3)/\surd 2$; 1 and $\pm i$.

18. $J_r = \{\ldots r-10, r-5, r, r+5, r+10, \ldots\}$
e.g. $J_1 = \{\ldots -9, -4, 1, 6, 11, \ldots\}$.

4.9 **2.** 1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1 elements; total 36 elements in A.

4. 8; 12. **7.** Rule 7(a). **8.** Rules 2(a), 4(a), 8(a).

10. 25, 40, 20, 20. Intersection of each of C', $A \cap C$, A' with each of Y, N, $(Y \cup N)'$.

23. Yes, accepting Axiom of Choice.

5.9 **1.** (i) high, some not rising; (ii) and (iii) not high, not all rising; (iv) high, all rising.

2. (a) $\sim p$; (b) $q \vee r$; (c) $\sim q \wedge r$; (d) $\sim p \wedge (\sim q \wedge r)$; ($e$) $p \wedge q \rightarrow p \wedge r$.

4. F F T
T T T
T F T

6.

		p	q	r	$\sim p \wedge q$	$q \vee r$	$q \rightarrow r$
Men	S	T	F	F	F	F	T
	M	T	T	T	F	T	T
	W or D	T	F	T	F	T	T
Women	S	F	F	F	F	F	T
	M	F	T	T	T	T	T
	W or D	F	F	T	F	T	T

18. a_1 and a_2 exhaustive: $P(a_1 \vee a_2) = 1$.

19. Evens.

21. Each $=4/20=1/5$.

23. Let $a_1 = B$'s hand better, $a_2 = B$'s hand worse, $a = B$ raises. Then: $P(a_1)=1/10$ and $P(a_2)=9/10$. $P(a \mid a_1)=9/10$ and $P(a \mid a_2)=1/5$. By Bayes: $P(a_1 \mid a) : P(a_2 \mid a) = 9 : 18$.

28. $\binom{n}{r} > \binom{n}{r-1}$ for $r = 1, 2, \ldots \frac{1}{2}(n-1)$; $\binom{n}{r} = \binom{n}{r-1}$ for $r = \frac{1}{2}(n+1)$.

6.9 **5.** For: $b * b^{-1} * a = e * a = a$; if $b * x_1 = b * x_2$ then $x_1 = x_2$.

9. If integers (excluding 0) are a group, then 0 never appears in $\times$ table.

16. Write rotation through 0° as 1. If rotations through 90°, 180°, 270° are κ, λ, μ, then $\lambda \times \lambda = 1$ and $\lambda = -1$; $\kappa \times \kappa = -1$ and $\kappa = i$; $\kappa \times \mu = 1$ and $\mu = -i$.

22. E.g. if $z_1 = ib$, $z_2 = jc$, then $z_1 z_2 = kbc$ but $z_2 z_1 = -kbc$.

30. Draw graphs of $y = x^2 + x + 1$ and $y = 1/(1+x^2)$.

7.9 **8.** Note: x is not brother of himself; y brother of x may imply x is sister of y.

9. For $xRy = x$ and y both even, xRx true if x even, false if x odd.

8.9 **3.** Gradient either AB/OA (1′ vertically for 12′ horizontally) or AB/OB (1′ vertically for 12′ up hill). Former is slope of OB.

10. First result: (x, y) and its negative (a, b) add to zero $(0, 0)$. So $(0, 0) = (x, y) + (a, b) = (x + a, y + b)$. $\therefore$ $a = -x$, $b = -y$.

13. No. $0 \leqslant \lambda \leqslant 1$ for points on PQ between P and Q; other λ for points not between P and Q.

19. Slope of P_1P_2, $m = (y_2 - y_1)/(x_2 - x_1)$ when $x_1 \neq x_2$. But m does not exist when $x_1 = x_2$ (line parallel to Oy); $m \to \infty$ as $x_1 \to x_2$ (Chap. 9).

32. $c^2 > r^2(a^2 + b^2)$; this is also condition that perpendicular distance from O to line is greater than r.

9.9 **4.** Inverse of $y = ax^2$ on $x \geqslant 0$ is $x = \sqrt{(y/a)}$ on $y \geqslant 0$ $(a > 0)$ but on $y \leqslant 0$ $(a < 0)$. Both increasing $(a > 0)$ or both decreasing $(a < 0)$.

8. Write $u = 1 - \sqrt{(1 + x)}$ defined on $x \geqslant -1$ with range $u \leqslant 1$.

19. $S_n \to \pm\infty$ if $r = +1$; S_n oscillates if $r = -1$.

24. (i) $y \to 0$, (ii) and (iii) $y \to 2$, (iv) $y \to \pm\infty$, (v) and (vi) $y \to \frac{1}{2}$. y is defined at $x = -1$ in (ii) and (v), not in (iii) and (vi).

29. $f(x)$ is $1/\sqrt{u}$ where $u = 1 + 1/x^2$.

10.9 **4.** Lines with slopes m and m' are perpendicular if $mm' = -1$; normal has slope $-1/f'(x_1)$. Tangent: $2x_1x - y - x_1^2 = 0$; normal:

$$x + 2x_1y - x_1(1 + 2x_1^2) = 0.$$

6. $D(\sqrt[q]{x^p}) = D(x^{p/q}) = \dfrac{p}{q}x^{(p/q)-1} = \dfrac{p}{q}\sqrt[q]{x^{p-q}}$.

12. $Dy = (1 + x^2)/(1 - x^2)^2$; $D^2y = 2x(3 + x^2)/(1 - x^2)^3$. $y \to \infty$ as $x \to 1$ from below.

13. Greatest height when $v = 0$, $t = u/g$, $x = u^2/2g$.

15. $R = px = 2x(1-x)$ with $\dfrac{dR}{dx} = 2(1-2x)$ (£ per item). R increases to $\frac{1}{2}$ as maximum (£500) at $x = \frac{1}{2}$ (500 items per week).

21. Integral is $-\frac{1}{2}\int u^{-2}\,du = \frac{1}{2}u^{-1}$. Consistent since $\frac{1}{2}\dfrac{1}{1-x^2} = \frac{1}{4}\left(\dfrac{1+x^2}{1-x^2}+1\right)$.

22. Write $\displaystyle\int x\frac{1}{\sqrt{x-1}}\,dx = x\int\frac{dx}{\sqrt{x-1}} - \int Dx\left(\int\frac{dx}{\sqrt{x-1}}\right)dx.$

23. $\int x^{m-1}(1-x)^{n-1}\,dx = \int(1-u)^{m-1}u^{n-1}(-1)\,du = -\int u^{n-1}(1-u)^{m-1}\,du$; limits (0, 1) for x go to (1, 0) for u, so that

$$\int_0^1 x^{m-1}(1-x)^{n-1}\,dx = -\int_1^0 u^{n-1}(1-u)^{m-1}\,du$$

$$= \int_0^1 u^{n-1}(1-u)^{m-1}\,du = \int_0^1 x^{n-1}(1-x)^{m-1}\,dx.$$

11.9 2. Yes, $x = 1 - \frac{2}{3}\sqrt{3}$.

9. Sufficient that $R'(\alpha) = C'(\alpha)$ and $R''(\alpha) < C''(\alpha)$.

10. Min. $z = 1$ at $x = 4/5$, $y = 1/5$.

15. Take sides of rectangle as $2x$ and $2y$; corners symmetrically on circle.

16. Alternative: smallest circle from which rectangle of given area can be cut.

17. Max. $R = px + qy$ subject to $\phi(x, y) = 0$ gives $\phi'_x : p = \phi'_y : q$ (by Ex. 13).

23. 3 or 4 terms of second series give $\log_{10} 2 = 0{\cdot}301$.

26. Proofs in Hardy, *Pure Mathematics* (10th Ed., 1952), Examples XXIX. Note: to avoid terminating decimals, express as recurring by Ex. 25.

27. First relation: write $e - S_q = \dfrac{1}{(q+1)!} + \dfrac{1}{(q+2)!} + \dots$ and multiply $q!$. Second relation: $\dfrac{p}{q}q! = p(q-1)!$ and $S_q q! = q! + q! + \dfrac{q!}{2!} + \dots$ (all positive integers). But there is no positive integer less than $1/q$.

29. α or β negative integer: series terminates; γ negative integer: series not defined after a finite number of terms.

30. Series: $1 + \frac{1}{2}x^2 + \frac{3}{8}x^4 + \frac{5}{16}x^6 + \dots$.

12.9 1. $De^{\pm\frac{1}{2}x^2} = \pm xe^{\pm\frac{1}{2}x^2}$.

9. Yes; y *decreases* at constant proportionate rate.

15. From $\int x_0 e^{rt}\,dt$, show $y = \dfrac{1}{r}x(1-e^{-50r})$ and $u = \dfrac{1}{r}x$.

21. $\int e^x \sin x\,dx = e^x \sin x - \int e^x \cos x\,dx$ and $\int e^x \cos x\,dx$ similarly; then substitute from one to other.

30. y symmetric about $x = 0$; so $\displaystyle\int_0^\infty y\,dx = \frac{1}{2}\int_{-\infty}^\infty y\,dx = \frac{1}{2}$. Then:

$$\int_0^\infty e^{-\frac{1}{2}x^2}\,dx = \tfrac{1}{2}\sqrt{2\pi} \quad\text{and}\quad \Gamma(\tfrac{1}{2}) = \sqrt{2}\int_0^\infty e^{-\frac{1}{2}t^2}\,dt = \sqrt{\pi}.$$

13.9 4. $x_2 = \frac{1}{A}\left\{ - y_1(a_{12}a_{33} - a_{13}a_{32}) + y_2(a_{11}a_{33} - a_{13}a_{31}) - y_3(a_{11}a_{32} - a_{12}a_{31})\right\}$

$x_3 = \frac{1}{A}\left\{y_1(a_{12}a_{23} - a_{13}a_{22}) - y_2(a_{11}a_{23} + a_{13}a_{21}) + y_3(a_{11}a_{22} - a_{12}a_{21})\right\}.$

5. First 2 equations (5) give x_1 and x_2 by (8) of 13.3. Substitute in third equation for relation:

$$y_1(a_{21}a_{32} - a_{22}a_{31}) - y_2(a_{11}a_{32} - a_{12}a_{31}) + y_3(a_{11}a_{22} - a_{12}a_{21}) = 0.$$

7. Inner products $\sum_{s=1}^{n} b_{rs}c_s$ for $r = 1, 2, \dots m$.

8. See (1) and (2) of 13.8.

13. (r, s)th element is $x_r y_s$ in xy' and $x_s y_r$ in yx'.

16. **AB** is $m \times k$ and **(AB)C** is $m \times n$.

19. If **A** is $m \times n$, first **O** is $n \times n$, second $m \times m$, third $m \times n$.

25. (r, s)th element is $\sum_{t=1}^{n} a_{rt}a_{st}$.

14.9 7. Particular integral $y = \frac{1}{1 - D^2}x = x$.

9. As $a \to \pm 1$, $y \to \infty$ for each x, i.e. oscillations of 'infinite' amplitude.

10. If $\alpha_{11}\alpha_{22} - \alpha_{12}\alpha_{21} \neq 0$, solve the equations for Dy and Dz.

12. Equations become $Dy = (a_{11} + ka_{12})y = (a_{22} + \frac{1}{k}a_{21})y$. Two values of k from: $a_{12}k^2 + (a_{11} - a_{22})k - a_{21} = 0$.

16. Solution: $y_n = Aa^n + B(-a)^n$ where $y_0 = A + B$ and $y_1 = (A - B)a$. Hence: $y_n = \frac{1}{2}y_0\{a^n + (-a)^n\} + \frac{1}{2a}y_1\{a^n - (-a)^n\}$, giving $y_n = y_0 a^n$ if n even, $y_n = y_1 a^{n-1}$ if n odd.

20. If S and C are the two Laplace Transforms, the results obtained give $S = \frac{1}{\alpha} - \frac{p}{\alpha}C$ and $C = \frac{p}{\alpha}S$, i.e. $S = \alpha/(p^2 + \alpha^2)$ and $C = p/(p^2 + \alpha^2)$.

INDEX

Numbers refer to pages and those in italics to exercises

PRINTED IN GREAT BRITAIN BY
ROBERT MACLEHOSE AND CO. LTD
THE UNIVERSITY PRESS, GLASGOW

DEFINITIONS AND NOTATIONS

References are to sections of chapters

$\sum$	summation (1.7)
$n!$	n factorial (1.7)
$\binom{n}{r}$	binomial coefficient (1.7)
$\lvert a \rvert$	absolute value or modulus (1.7)
R	field of rational numbers (2.1, 2.2)
$R(\surd 2)$	adjunction, of $\surd 2$ to R (2.3)
R^*	field of real numbers (2.4)
GLB	greatest lower bound (LUB similarly) (2.4)
i	complex unit, $i^2=-1$ (2.5)
z	complex number $x+iy=r(\cos\theta+i\sin\theta)=re^{i\theta}$ (2.5, 12.7)
r, θ	absolute value and argument, of z (2.5)
C	field of complex numbers (2.5)
J^+	set of positive integers (natural numbers) (2.6)
J	integral domain of integers (2.6)
mod n	modulo n (2.7)
$F[x]$	polynomials over field F (3.3)
$F(x)$	rational fractions, adjunction of x to F (3.4)
ω	nth root of unity (3.8)
$\{a \mid a \text{ is } P\}$	set (4.1)
$\in$	belongs to (4.1)
$\subset$	proper subset of (4.1)
A'	complement, of set A (4.2)
$\cap, \cup$	intersection and union, of sets (4.2)
U, ϕ	universal and empty sets (4.2)
d, c	infinite cardinal numbers (4.7, 4.8)
$\sim p$	negation of p (not) (5.1)
$\wedge, \vee$	conjunction and disjunction (and, or) (5.1)
$\rightarrow$	implication, many-one mapping (5.1, 7.3)
$\leftrightarrow$	equivalence, one-one mapping (5.1, 7.3)
$P(a)$	probability, of statement a (5.5)
$P(a_1 \mid a_2)$	conditional probability, of a_1 given a_2 (5.6)
G	group, with identity e and a^{-1} inverse of a (6.2)
$r(A)$	operator; $sr(A)$ successive operators (r first, s second) (6.4)
F	field, with identities 0 and 1 (6.5)
$X . Y$	Cartesian product, of sets (7.1)
yRx	statement of a relation R (7.1)
$y=f(x)$	function (7.3)

$X \xrightarrow[f]{} Y$	mapping; $X \xrightarrow[T]{} Y$ transformation (7.3)
$\cong$	isomorphic (7.4)
$Z=f(z)$	function of a complex variable (7.6)
V	vector space (8.3)
$V_n(F)$	space of n-tuples over F (8.4)
$E_n(F)$	n-dimensional Euclidean space over F (8.4)
$(ABCD)$	cross-ratio $\dfrac{AB\,.\,CD}{AD\,.\,CB}$ (8.7)
$(1, \pm i, 0)$	circular points at infinity (8.8)
$[a, b]$	interval $a \leq x \leq b$ (9.3)
N	neighbourhood, of α (9.3)
$f^{-1}(x)$	inverse function (9.3)
$\operatorname*{Lim}_{n\to\infty} f(n)$	limit of $f(n)$ as n increases without bound (9.5)
$\operatorname*{Lim}_{x\to\alpha} f(x)$	limit of $f(x)$ as x approaches α (9.6)
$y', f'(x)$ $Dy, Df(x)$ $\dfrac{dy}{dx}, \dfrac{d}{dx}f(x)$	derivative of $y=f(x)$ (10.2)
$\int_a^b f(x)\,dx$	definite integral of $f(x)$, area (10.5)
$\int f(x)\,dx$ $D^{-1}f(x)$	indefinite integral of $f(x)$, anti-derivative (10.6, 10.8)
$D^n f(x)$	nth derivative, ($-n$)th integral, of $f(x)$ (10.8)
$\operatorname{Max} f(x)$	local maximum of $f(x)$ (minimum similarly) (11.2)
$\sum u_n$	infinite series (11.3)
$\sum a_n x^n$	power series, with radius of convergence r (11.6)
$\pi = 3{\cdot}14159\ldots$	Archimedes' constant (12.1, 12.5)
$e = 2{\cdot}71828\ldots$	Euler's constant, $\operatorname*{Lim}_{n\to\infty}\left(1+\dfrac{1}{n}\right)^n$ (12.1, 12.2)
e^x, $\exp x$	exponential function (12.2)
$\log x$	logarithmic function (12.3)
a^x, x^a	power functions (12.4)
$\cos x, \sin x$ $\tan x, \tan^{-1} x$	circular functions (12.5), trigonometric functions (12.7)
$\cosh x, \sinh x$	hyperbolic functions (12.6)
$\mathbf{A} = \lVert a_{rs} \rVert$	matrix (13.4)
$\mathbf{A}'$	transpose, of matrix (13.5)
$A = \lvert \mathbf{A} \rvert$	determinant, of matrix (13.6)
$\mathbf{A}^{-1}$	inverse, of matrix (13.6)
$\bar{y}(p)$	Laplace transform, of $y(t)$ (14.7)
$\bar{Y}(s)$	generating function, of Y_n (14.7)

BASIC MATHEMATICS

R.G.D. ALLEN

MACMILLAN